Correcting for the Dependence Structure in Social Networks

by
Caroline Epstein

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, Maryland
April, 2014

Abstract

The use of social network data has recently become increasingly prevalent in social science research and in clinical fields. While some researchers deliberately exploit the social network structures to maximize response rates, reach hidden populations, or learn about the transmission of information or diseases from one person to another through network connections, others unintentionally sample observations from connected networks within the overall target population; the latter is especially common when samples are collected from contiguous geographic areas or similar institutions. Statistical inference from observations sampled from social networks is problematic because the observations are often inherently correlated, but this dependence is rarely adequately accounted for in statistical inference. Failing to account for the dependence between network observations has unfavorable, sometimes dangerous, consequences for inference, causing underestimated standard errors, inflated statistical significance, and high type I error rates. Throughout this work, we demonstrate the gravity of these repercussions through simulations, which entail constructing network structures resembling realistic social networks, associating independent outcomes with each subject, and generating various levels of dependence in the sample. We sample the generated outcome data to draw inferences about the population mean, incorrectly assuming independence between observations. We find that ignoring network dependence has devastating consequences for the validity of inference, and become more severe with increasing correlation: estimated coverage of 95% confidence intervals dropped as low as 33% when the sample exhibited high dependence. We suggest informal methods for quantifying and accounting for dependence in various research settings, but each with the objective of drawing valid inferences for a population mean. We demonstrate the efficacy of these methods by implementing them in all simulated dependence settings. We found that by employing these methods, we were able to attain valid, or nearly valid, inference which we assess through estimated cov-

erage. An important objective of future work in this area is to extend these methods to allow for more general applications.

Author: Caroline Epstein

Readers: Dr. Elizabeth Ogburn and Dr. Jeffrey Leek

Preface and Acknowledgements

I am more grateful than I can express to my advisor, Dr. Elizabeth Ogburn, for her patience, guidance, and dedication to being an accessible mentor, even through her growing responsibilities in the department. I would like to thank her for helping me navigate my way through this research project; I greatly appreciate both the freedom she gave me to work and problem-solve independently, and her invaluable advice when I would struggle for direction.

I am incredibly appreciative of the support from my professor and thesis committee member, Dr. Jeffrey Leek. He taught me not only a wealth of statistical analysis procedures, but how to think critically about the research objective at hand and determine the most efficient method for addressing it. I would specifically like to thank him for the thoughtful and constructive advice he offered on this project.

Lastly, I would like to thank all of my colleagues and professors in the Biostatistics department for all that you have taught me, and the friendly and supportive environment that you have helped foster.

Contents

List of Tables

List of Figures

1 Introduction

Interest in and use of social network data has grown rapidly in recent years, for research in the social sciences and in many different clinical settings (e.g. [5] [10] [17] [19]). The statistical analysis of observations sampled from a social network is still in its nascency, and methods for valid inference for many types of network data have yet to be developed. Statistical inference using observations from a social network is problematic because these observations are often inherently (and positively) correlated, yet many researchers frequently fail to account for this dependence in their analyses. Sometimes this is because adequate methods for accounting for dependence are not yet available, but often it is because researchers are unaware of the dependence among observations, and perhaps even unaware of the underlying network structure linking the sampled subjects.

Incorrectly treating network observations as independent is potentially dangerous because it results in underestimated standard errors, inflated statistical significance, and anticonservative inference. Throughout this work, we will illustrate these consequences via simulations that depict a very simple research setting: we are interested in estimating the population mean, μ, of some outcome of interest, Y, from a set of observations sampled from subjects who are members of a network. We assume that the observations are identically distributed, but we allow them to be non-independent. From the observed network data, we use the sample mean, $\bar{Y}$, the estimated standard error of $\bar{Y}$, and the assumption that $\bar{Y}$ converges to a normal distribution to draw inferences about μ. While the sample mean remains unbiased for μ with dependent data, the variance of $\bar{Y}$ increases when observations are correlated, and we will therefore underestimate the standard error of $\bar{Y}$ if we incorrectly treat the observations as independent. Underestimating the standard error undermines the validity of inference, leading to anti-conservative p-values, high type I error, and low coverage

probabilities.

We derive an analytic expression for the factor by which we underestimate the standard error of $\bar{Y}$ and quantify the effect of this underestimation on inference in the simple research setting described above, using data simulated with various levels of correlation between observations. We propose methods to account for dependence in the estimation of standard error, and apply these methods to simulated data to demonstrate the possibility of recovering valid or almost valid inference using dependent network data. Though the setting we consider is very simple, accurately estimating the standard error of a sample mean is central to a wide range of interferential methods and the results we present are easily extended to most M-estimation approaches to statistical inference.

In Section 1.1 we will discuss areas of research in which researchers fail to account for dependence due to underlying network structure when drawing inferences. In Section 1.2, we will introduce common network topology terms that we use throughout the paper, and briefly discuss some simple network generating models. We modify one such model to construct the network structures used in the simulations on which we base many of our results; these simulations, described in Section 2, entail building a realistic social network structure, associating independent outcome values with subjects in the network, and generating dependence between observations. In Sections 3 -3.4.2, we will demonstrate that failing to account for network dependence compromises the validity of statistical inference; specifically we illustrate, via simulations, the type I error rate when treating network observations as independent under various dependence settings. In Section 4, we suggest informal methods for quantifying and accounting for dependence in network samples and demonstrate their performance by implementing them in our simulations. If a significant amount of information

is known about the network generating process and the source of dependence, then perhaps other, more specific methods can be used to incorporate such information. However, we believe that this is rarely the case, and the methods that we suggest intend to address the issue of dependence in inference when little to nothing is known about the source. These methods represent an important first step towards controlling for dependence in statistical inference using observations sample from subjects in a network. They are somewhat ad hoc and may not be generalizable to more complex settings than the ones we consider here; an important objective of future work in this area is to extend these methods to allow for more general applications.

1.1 Examples from the Literature

Respondent-driven sampling (RDS), a variant of snowball sampling introduced in 1997 [1], is widely used today to reach hidden populations such as injection drug users, groups at risk for HIV, gang members, and other socially stigmatized or hard to reach groups. Standard probability sampling in these populations tends to result in low response rates, but RDS exploits the social networks in such populations by drawing subsequent participants into the sample through relationships with current sample members and incentives (e.g. food coupons) both to participate and to recruit [1]. Researchers acknowledge that RDS estimators are sometimes sensitive to the initial subjects sampled and that preferential referral behavior can result in biased estimates [3] [4]. Some have developed methods to counter these issues [2] [5], however in general the literature on RDS does not acknowledge or account for dependence due to the underlying network structure. Analyses of RDS samples that fail to account for network dependence will fall prey to the problems of deflated standard error estimates and inflated measures of significance that we describe below, and the conclusions should therefore be read with caution. It has been demonstrated that underlying network structure does affect the performance of the RDS estimator [6],

and very recently methods have been proposed to account for such dependence [7]; however, we are unaware of these methods having been used in practice, and they require modeling the dependence structure explicitly, which may not always be feasible. The methods that we propose for quantifying network dependence do not require parametric assumptions about the dependence structure.

Social networks are frequently used to study peer effects, that is, the causal effect of one individual's outcome on the outcomes of his or her peers. Christakis and Fowler published a series of high-profile papers [10] [11] [12] purporting to find significant peer effects for outcomes such as obesity, smoking, and happiness through analysis of a large social network (the Framingham Heart Study). This work has inspired numerous research programs that study peer effects using the same statistical methods [20] [21] [22] [23]. However, these methods have come under considerable criticism [13] [24] [25] [26]; while some of the problems have been addressed in subsequent analyses by Christakis, Fowler, and others, none of these analyses adequately account for the social network dependence inherent in the observations. Since incorrectly treating network observations as independent leads to anti-conservative p-values and high type I error, this should cause skepticism about the significance of the findings. Driven by the conjecture of significant peer effects, social networks have also been used to study the efficacy of network-oriented interventions for risky behaviors [14] [15] [27] [28]. These studies similarly fail to account for the dependence between observations in the networks studied when drawing inferences, and are therefore likely to suffer from inflated statistical significance which could call into question the true effectiveness of such interventions.

Studies of infectious diseases frequently use social networks to learn about the social aspect of disease spread (e.g. contact patterns between subjects in a network),

or about structural characteristics of networks that contribute to the spread of infectious diseases (e.g. the network-based distance between subjects) [18] [19]; both of these objectives require accounting for dependence due to network structure to allow for valid statistical inference, yet many researchers fail to do so [18] [19]. Oftentimes, researchers studying infectious disease patterns devise sampling schemes which they sincerely believe will yield independent observations. However, when the sample is collected over a contiguous geographic area, or when the disease has become a pandemic outbreak, truly random samples are generally not possible since all observations are most likely sampled from the same transmission network.

1.2 Network Topology and Generating Models

A social network is comprised of a set of individuals, called nodes, and pairwise relationships between them, called edges or ties[8]. While networks can be defined with multiple types of relationships possible between pairs of subjects [8], we will only consider binary ties indicating the presence or absence of a relationship. That is, we define a random variable T_{ij} to be equal to 1 if a tie exists between subjects i and j, and $T_{ij} = 0$ otherwise, for $i, j \in (1, .., n)$ and n equal to the number of nodes in the network. Ties can be directed, where $T_{ij} = 1$ does not imply that $T_{ji} = 1$, or they can be undirected, in which case all ties are reciprocated [8]. In our work, for the purpose of simplicity and because they are arguably more common in the literature on social networks, we will only consider undirected networks (networks in which all ties are undirected).

A common summary measure of network structure is the density of the network, defined as the proportion of all possible edges that are realized in the network [8].The

density of graph G, comprised of n total nodes, is given by

$$dens(G) = \frac{\sum_{i=1}^{n} \sum_{j=i}^{n} T_{ij}}{\binom{n}{2}}.$$

Holding all other factors constant, dependence usually tends to increase with the density of a social network. The degree of node i is defined as the total number of ties that node i has to other nodes in the network [8]:

$$deg(i) = \sum_{j=1}^{n} T_{ij}.$$

The distance, or degree of separation, between nodes i and j is defined as the number of edges in the shortest path between the two nodes. For example, if $T_{ij} = 1, T_{jk} = 1$, and $T_{ik} = 0$ (nodes i and j share a tie, nodes j and k share a tie, but nodes i and k do not share a tie), then the distance between nodes i and k is 2. Network dependence is, roughly, dependence that tends to result in larger correlations between node attributes for nodes that are closer in network distance. We say that the network forms one connected component if there exists a sequence of pairwise ties allowing a path between any two nodes in the network.

Many parametric models exist to generate social network structures. One of the most straightforward network generating models is the Erdös-Rényi model, where T_{ij} is randomly drawn from a Bernoulli distribution with a fixed parameter, p, representing the probability of an edge between any pair (i, j) [9]. While this model has the advantage of simplicity, it is generally not representative of realistic social networks since it assumes that all edges are independent and equally likely. Slightly more appropriate for social networks is the Barabasi-Albert model, which allows for preferential attachment – stipulating that the probability of a new node i forming a tie with an existing network node j increases with the degree of node j [9] – to

generate scale-free networks. Scale-free networks are networks with a degree distribution that follows a power law, meaning that the proportion of nodes of degree k, $P(k) \propto \frac{1}{k^r}$, where usually, $r \in [2, 3]$; this tends to result in a large number of nodes of low degree and a small number of nodes with very high degree. There is strong evidence to suggest that social networks tend to be scale-free [8] .

One of the most commonly used and studied models for social network analysis, however, is the latent space model [29]. According to the latent space model paradigm, individuals who have similar characteristics are close in "social space", and the probability of a tie forming between two individuals is modeled as a function of these social distances [9]. For example, if we let X represent a continuous latent random variable, then a pair of subjects i and j having similar values, X_i and X_j, will be close in social space, resulting in a relatively high value of $P(T_{ij} = 1)$, the probability of a relationship forming. All ties are assumed to be independent [9] conditional on locations in social space. By letting $\mathbf{A}$ be the $n \times n$ matrix summarizing the relationships between the n subjects in a network (the $[i, j]$ entry corresponding to the presence or absence of a tie between i and j, T_{ij}), we use conditional independence to model the likelihood of the set of relationships in realized networks, $\mathbf{A}$:

$$P(\mathbf{A}|X, C, \theta) = \prod_{i \neq j} P(T_{ij}|X_i, X_j, C_{ij}, \theta),$$

where C and C_{ij} are other optional observed characteristics, perhaps pair-specific, and θ a population parameter contributing to the probability of ties [29]. We can

parameterize the above expression as a logistic regression model where,

$$\begin{aligned}\eta_{ij} &= logodds(T_{ij} = 1|X_i, X_j, C_{ij}, \alpha, \beta) \\ &= \alpha + \beta' C_{ij} - |X_i - X_j| \\ P(T_{ij}) &= \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}.\end{aligned}$$

The above expression implies that for any two subjects i and j that are equidistant in social space from subject k, the log odds ratio of $T_{ik} = 1$ to $T_{jk} = 1$ is given by $\beta'(C_{ik} - C_{jk})$. We construct the social networks for the simulations presented in this paper with a latent space approach due to their flexibility and ability to reasonably emulate realistic social networks, slightly simplifying the model above for determining the probability of a tie. Namely, we set $\alpha = \beta = 0$ (we do not consider other observed characteristics), and model the probability of a tie between any two subjects i and j simply as a function of their positions in social space. In the following section, we will explicitly describe how our latent space networks are constructed and then explain how we generate dependence between nodal attributes in these networks.

2 Simulation Settings: Constructing Networks and Generating Dependence

Most of the results we present are based on simulations of scenarios in which network dependence poses a challenge for statistical inference. Below we describe the latent space model that we use to generate the networks in all of our simulations. The latent space model generates a network topology, i.e., a list of edges between pairs of nodes. Once we have simulated the network we simulate an outcome, Y, for each node in the network. We discuss two different methods for generating dependence in the sample $Y_1, ..., Y_n$ of outcomes associated with nodes in the in the network. Results from

both dependence-generating methods will be presented and compared throughout; although results are presented separately for the two dependence-generating settings, the methods that we use to quantify and correct for dependence are identical and we do not assume any knowledge of the source of dependence. Simulations allow us to accurately estimate the distribution of an outcome in a network. Given a single complex network (even one which we have simulated ourselves), the dependence structure of the outcome variable (and therefore its variance) may be very difficult to estimate; these parameters estimated over a large number of simulations will serve as the gold standard to which we will compare our methods in later sections.

2.1 Constructing networks using a latent space model

We use a latent space model to simulate social networks comprised of n nodes, or subjects, each. We associate a continuous latent variable, X, with each node in the network. The probability of an edge between any two nodes is determined by the difference in their X values. This variable could represent income, education level, a measure of geographic location, genetic factors, or any characteristic that could reasonably contribute to the probability of two subjects having a relationship. For simplicity, we let X follow a normal distribution with mean 0 and variance 1, and we independently assign a value of this covariate to each of the n subjects in the network. That is, we generate $X_1, X_2, ..., X_n \overset{i.i.d.}{\sim} N(0,1)$, and each subject, i, is assigned to have covariate X_i.

We determine the probability of a tie forming between any two subjects, i and j, by the similarity of their covariate values, X_i and X_j. We set the probability of a tie forming between i and j, $P(T_{ij})$, to be large when the absolute difference between X_i

and X_j is small, and vice versa. To achieve this, we let

$$\log\left(\frac{P(T_{ij})}{1-P(T_{ij})} \;\middle|\; X_i, X_j\right) = f(|X_i - X_j|),$$

where $f(|X_i - X_j|)$ is large and positive when $|X_i - X_j|$ is small, near zero when $|X_i - X_j|$ is moderate, and large and negative when $|X_i - X_j|$ is large. We split the absolute differences into two categories, *similar* and *dissimilar*, and define f separately for each category. For similar pairs, that is $|X_i - X_j| < 0.05$, we define $f_1(|X_i - X_j|)$ to be the inverse of the absolute difference, and for dissimilar pairs, that is $|X_i - X_j| > 0.05$, we define $f_2(|X_i - X_j|) = -5 \cdot |X_i - X_j|$. The threshold of 0.05 differentiating similar from dissimilar latent variable values is specific to the distribution we chose for X, and almost always guarantees that subjects with very similar latent variable values will form a relationship while ensuring that subjects with increasingly dissimilar latent variable values will have decreasing probabilities of ties forming.

Using the functions defined above, we set the probability of an edge between subject i and subject j to be

$$P(T_{ij}) = \frac{\exp\{f(|X_i - X_j|\}}{1 + \exp\{f(|X_i - X_j|\}}.$$

We defined $P(T_{ii})$ to be zero for all i. The presence or absence of a tie between subjects i and j is then determined by generating a new random variable, T_{ij}, where $T_{ij} \sim Bernoulli(P(T_{ij}))$. If $T_{ij} = 1$, we form an undirected tie between subjects i and j, and if not, they remain unconnected. We will refer to all pairs of subjects who share a tie as "neighbors" or "friends." Our specifications for f_1 and f_2 entail that the overall density of the graph is usually great enough to ensure that population of nodes form one connected component. However, in the occasional cases where $c > 1$

components formed, we added a total of $(c-1)$ additional edges between randomly selected nodes in the various components in order to fully connect the network. At this point, the network is fully constructed; we now turn to simulating an outcome Y with dependence on top of the network structure.

2.2 Peer influence: dependence generated over time

In the first of the two settings we consider throughout, dependence mimics influence, information, or a contagious process traveling through the network from node to node. All subjects' outcomes are initially generated as independent random variables. We then simulate the subjects interacting with one another according to their network ties, at discrete time points, and at each time point each subject's outcome is influenced by his neighbors' outcomes, resulting in increasing dependence among the outcomes over time.

Let Y_i^t represent the outcome for subject i at time t. We let Y^0 follow a normal distribution with a mean of zero and variance of one. That is, we randomly generate n outcomes where

$$Y_1^0, Y_2^0, ..., Y_n^0 \overset{i.i.d.}{\sim} N(0,1)$$

and each subject i receives an initial outcome value of Y_i^0. For time $t > 0$, the outcome for subject i at time t is given by a weighted average of his own outcome and the outcomes of all of of his neighbors at time $t-1$. Let v_i^t represent the average of all of subject i's neighbor's outcomes at time t; explicitly,

$$v_i^t = \frac{1}{deg(i)} \sum_{j=1}^{n} T_{ij} \cdot Y_j^t.$$

We determine the weight given to v_i^t by a random susceptibility probability, p_i^t, which indicates how susceptible subject i is to his neighbors' influence at time t. At each time point and for each subject, we generate a new susceptibility probability $p_i^t \sim Uniform[0, m]$, where $m < 1$ is the maximum susceptibility probability that any subject could have. Explicitly, at time t, we calculate subject i's outcome as

$$Y_i^t = (1 - p_i^t) \cdot Y_i^{t-1} + p_i^t \cdot v_i^{t-1}.$$

The amount of dependence in the resulting network depends on three features of the simulation setting: how large we set m, the maximum susceptibility probability; the number of time points at which the subjects interact; and the density of the network. At each time point the outcomes become increasingly similar, eventually converging to a common value when the number of time points becomes sufficiently large (this is true for undirected networks but not necessarily for directed ones). The rate of convergence to this common value is determined in part by the value of m, and in part by the density of the network.

For all simulations we set n = 100. We fixed m = 0.08 throughout (meaning that a maximum of 8% of a subject's outcome at time t will be determined by friends' previous outcomes) and fixed the network topology across simulations, varying only the number of time points to generate different levels of dependence. Fixing the number of time points and varying either the maximum susceptibility threshold or the density of the graph would have led to similar results. We analyzed the distribution of the outcome at times t =0 (corresponding to a setting where observations are truly independent), 30, 60, and 90. In what follows we refer to this simulation setting as the "peer influence" setting.

2.3 Latent variable dependence: outcomes correlated with latent space variable

An alternative source of dependence among outcomes $Y_1, ..., Y_n$ sampled from nodes in a social network is correlation between the outcome and a latent variable that predicts graph structure. This setting differs than the previous one in that dependence is only realized through a sampling method that depends on edges in the network, which are inherently correlated with the outcome of interest. Oftentimes, researchers collect observations that they presume to be independent, but when the observations are sampled from a small geographic area (e.g. a single neighborhood or town), or from a single or similar institutions (e.g. professors in academia), the sample will often, if unintentionally, represent a sub-network from the overall target population and network dependence may be present in the sample. Latent variable dependence could also be present when sampling methods deliberately exploit the relationships in networks to maximize response rates through chain-referral methods, such as respondent driven sampling or snowball sampling.

As an example, consider a network of families in a U.S. city. Suppose that the latent variable, X, according to which we generated the graph topology in Section 2.1, represents parenting philosophy, so that families who share the same parenting philosophy are more likely to be tied in the network than are families with different parenting philosophies. Now suppose that researchers are interested in studying some behavioral outcome for children in daycare, and that they collect a sample of children from one or a small number of daycare facilities in the city. If families that are associated with one another in the network are more likely to send their children to the same daycare, then it could be the case that a sample ascertained in this way is comprised of dependent rather than independent observations. Even if these observations provide unbiased estimates of the estimands of interest to the researchers (i.e.

there is no systematic selection bias), valid inference is not possible without taking the dependence in the sample into account. When our sampling method depends on the connections between subjects, and these connections are correlated with the outcome of interest, the observations in our sample will generally be dependent. We could collect independent observations by conducting a random sample from all parts of the income distribution, but unfortunately, entirely random sampling is not always a feasible option.

In order to generate latent variable dependence, we modify the network generating procedure described in Section 2.1 and generate a latent variable - outcome pair (X, Y) for each subject, drawn randomly from a bivariate normal distribution, where each marginal distribution is a $N(0, 1)$ and there is some positive correlation between the latent variable, X, and the outcome, Y. For each simulated network we generate 200 pairs

$$(X_1, Y_1), (X_2, Y_2), ..., (X_{200}, Y_{200}) \overset{i.i.d.}{\sim} BVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

with $\rho > 0$. We construct the network based on the latent variable, exactly as previously described, and from this total population of 200 we will select a subsample of $n = 100$ observed subjects.

We use snowball sampling to select $n = 100$ observations, but any sampling scheme that gives priority to nodes that are close in network distance would result in a dependent sample. We first select the outcome for the most highly connected subject in the underlying network of 200 nodes; in the case of a tie, we give preference to the node that we generated first in the network (e.g. the minimum i where $i \in [1, 200]$). We build our sample by selecting the outcomes of all of the neighbors of the initial node,

then the outcomes of the neighbors of these neighbors, and so on, working outwards until we have reached our target sample size of 100 observations.

The amount of dependence in our resulting 100-node "observed" network is determined by ρ, the correlation between X and Y, and by the density of the underlying 200-node network. We fix the density by fixing similarity category functions, f_1 and f_2 (defined in Section 2.1), and control the degree of dependence desired in the snowball sample by modifying the value of ρ. Below we consider the distribution of the outcome when $\rho = 0$ (corresponding to a sample of independent outcomes), $0.25, 0.55$, and 0.85. Throughout we refer to this simulation setting as the "latent variable dependence" setting.

In the next section, we demonstrate the consequences of failing to account for network dependence when making statistical inference about the true mean of Y using data simulated under these two different dependence settings. We will show analytically that the negative consequences for inference about μ using dependent data arise from biased estimation procedures for the variance of $\bar{Y}$ which result in underestimated standard errors. The sample mean, on the other hand, remains unbiased for μ even when the data is highly correlated. Consider the dependent outcomes generated by latent variable dependence, described above. The marginal distribution of any observation collected by the snowball sample is $Y_i \sim N(0, 1)$. Because the expectation operator is linear (even with dependence), it immediately follows that

$$E[\bar{Y}] = \frac{1}{n} \sum_{i=1}^{n} E[Y_i] = \mu = 0.$$

In the peer influence setting (described above in Section 2.2), the marginal distribution of any observation, Y_i is initially distributed as i.i.d. $N(0, 1)$, and therefore, at time

0, $E[\bar{Y}] = E[Y_i] = \mu = 0$ by the same argument as above. However, the sample mean also provides an unbiased estimate for μ after any given number of rounds of interaction. At time 1, $Y_i^1 = (1 - p_i^1)Y_i^0 + p_i^1\,(v_i^0)$. The quantity v_i^0 (defined in 2.2), representing the average of all of subject i's neighbors' outcomes at time 0, is clearly unbiased for $\mu = 0$ (as it is the sample average of i.i.d. variables from a $N(0, 1)$ distribution). Then,

$$E[Y_i^1] = (1 - p_i^1)E[Y_i^0] + p_i^1\left(E[v_i^0]\right) = \mu = 0,$$

and by the linearity of expectation, $E[\bar{Y}^1] = \mu = 0$. Iterating this argument, it is clear that the same result holds for any $t \geq 0$, and therefore $E[\bar{Y}^t]$ is unbiased for μ.

We will now demonstrate the consequences of failing to account for the dependence in correlated samples for inference about μ by analytically identifying the cause of bias in estimating the variance of $\bar{Y}$. We also illustrate the impact on coverage probabilities using the simulated data described in the sections above.

3 The consequences of ignoring dependence among observations sampled from a network

Recall that we would like to estimate the population mean, μ, of some outcome of interest, Y, from a sample of n identically distributed but correlated observations sampled from nodes in a network. In this section we investigate what happens when inference about μ makes use of the standard error of $\bar{Y}$ estimated under the assumption of independence. We will use $\bar{Y}$ as an estimator of μ throughout, and as we demonstrated above this estimator is unbiased under dependence or under in-

dependence. We also assume that $\bar{Y}$ is approximately normally distributed under dependence as it is under independence. (For discussion of this latter assumption see Section 4.1)

Throughout, we assume that dependence is due to positive correlation. (This is generally what we would expect to manifest in social network contexts, with either latent variable dependence or peer influence. Strictly negative correlation between pairs of friends is not consistent with most network topologies.) In most cases, the most egregious problem with treating observations as independent when they are not is that standard error of the sample mean is formulated differently for dependent outcomes than it is for independent outcomes. The estimated standard error of $\bar{Y}$ assuming independence will generally underestimate the true standard error when the sample $Y_1, .., Y_n$ exhibits dependence. Dependence among observations means that each individual observation contributes less information about the population mean than an independent observation from the same marginal distribution would contribute. Therefore, this problem can be understood as an issue of sample size: a sample of n dependent observations will produce the same inference for μ as some $n_e < n$ independent observations with the same marginal distribution.

Even if we account for the dependence among observed outcomes in our formulation of the standard error, we will still run into a second problem, namely estimation of the marginal variance σ^2 of Y . Estimation of the standard error requires knowledge of the variance of the outcome, which can be estimated with the sample variance, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$ when observations $Z_1, ..., Z_n$ are independent. However, when the observations are dependent, s^2 is not an unbiased estimator of σ^2; it will generally underestimate σ^2. For most realistic data generating mechanisms this problem is negligible for large n, but in small samples it could result in significant bias. We

will investigate these two issues analytically in this section, and then illustrate their impact with simulated data in Section 3.4.

3.1 Standard error estimation

To understand the differences between standard error estimation assuming independence versus accounting for dependence, we will consider the two settings separately, beginning with estimation under independence. Suppose that we have collected n observations, $Z_1, Z_2, ...Z_n$, which are independent and identically distributed, each with marginal mean μ and variance σ^2. Let $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, the sample mean. Then we know that $E[\bar{Z}] \to \mu$ with a rate of convergence that is determined by $var(\bar{Z})$. We can calculate the variance and standard error of $\bar{Z}$:

$$\begin{aligned} var(\bar{Z}) &= var\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) \\ &= \frac{1}{n^2} var\left(\sum_{i=1}^{n} Z_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} var(Z_i) \quad \text{(due to independence)} \\ &= \frac{\sigma^2}{n} \\ SE_{\bar{Z}} &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Then, by the Central Limit Theorem, as $n \to \infty$, $\sqrt{n}\left(\bar{Z} - \mu\right) \xrightarrow{d} N(0, \sigma^2)$. This is what justifies the approximation $\bar{Z} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ on which inference about $\bar{Z}$ is often based.

Now suppose that we have n observations, $Y_1, Y_2, ..., Y_n$, which are identically distributed with marginal mean μ and variance σ^2, just like Z, but are not independent. We assume here that a central limit theorem still holds for $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. (Un-

derstanding when this is reasonable for social network data is an important topic of research but is beyond the scope of this paper; the results of our simulations demonstrate that a CLT holds for a wide range of data generating mechanisms and degrees of dependence, including all of the settings with dependence that we consider (see Section 4.1). Then $E[\bar{Y}] \to \mu$ with a rate of convergence that is again determined by $var(\bar{Y})$. Define $b = \frac{1}{n}\sum_{i \neq j}^{n} cov(Y_i, Y_j)$. We can express the variance and standard error of $\bar{Y}$ in terms of b:

$$\begin{aligned}
var(\bar{Y}) &= var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{n^2} var\left(\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n} var(Y_i) + \sum_{i \neq j}^{n} cov(Y_i, Y_j)\right] \\
&= \frac{\sigma^2 + b}{n} \\
&= \frac{\sigma^2}{n}\left(1 + \frac{b}{\sigma^2}\right) \\
SE_{\bar{Y}} &= \frac{\sigma}{\sqrt{n}}\sqrt{\left(1 + \frac{b}{\sigma^2}\right)}
\end{aligned}$$

Therefore, by the Central Limit Theorem, as $n \to \infty$, $\sqrt{\frac{n}{1 + \frac{b}{\sigma^2}}}\left(\bar{Y} - \mu\right) \xrightarrow{d} N(0, \sigma^2)$. This result justifies the approximation $\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\left(1 + \frac{b}{\sigma^2}\right)\right)$ in order to draw inferences about μ using $\bar{Y}$.

We defined b as the sum of all pairwise covariances ($\binom{n}{2}$ terms) divided by the number of observations (n). The magnitude of b is not inherently meaningful, but we can compare it to two extreme values to get a sense of the amount of dependence in the network. When the observations are independent, b will be zero. At the other

end of the spectrum, when the observations are perfectly correlated, $b = (n-1)\sigma^2$. In most settings with network dependence, b will fall somewhere between these two values; whether and how close it is to each endpoint gives a qualitative measure of how much dependence is in the sample of outcomes.

Because b is non-negative when the expected correlation between observations is non-negative, the standard error $SE_{\bar{Y}}$ for dependent observations is larger than the standard error for the equivalent sample of independent observations by a factor of $\sqrt{(1+\frac{b}{\sigma^2})}$. As the amount of dependence in the sample $Y_1, Y_2, ..., Y_n$ increases, so does the ratio of b to σ^2, and, therefore, the discrepancy increases between the true standard error and the standard error calculated under the assumption that observations are independent. When $\frac{b}{\sigma^2}$ is large, we will grossly underestimate the standard error of $\bar{Y}$ when assuming independence, which leads to anti-conservative p-values and low coverage probabilities.

3.2 Effective Sample Size

One way to understand the discrepancy between $SE_{\bar{Z}}$ and $SE_{\bar{Y}}$, despite the fact that Z and Y have the same marginal distributions, is as an issue of "effective sample size." A sample of n positively correlated observations provides less information about the population mean, μ, than a sample of n independent observations drawn from the same marginal distribution. More generally, a sample of n positively correlated observations provides the same amount of information about μ as some $n_e < n$ independent observations drawn from a population with the same marginal distribution, where n_e is the effective sample size. This is obvious when the dependent observations are perfectly correlated: a sample of n perfectly correlated observations is equivalent to and provides the same amount of information about μ as a sample of just one independent observation.

Informally, the effective sample size is the number of independent observations that result in the same variance for the sample mean as n dependent observations. Using the expression for $var(\bar{Y})$ derived in the previous section, we solve for the effective sample size as follows:

$$var(\bar{Y}) = \frac{\sigma^2 + b}{n} = \frac{\sigma^2}{n_e}$$
$$n_e = n \cdot \left(\frac{\sigma^2}{\sigma^2 + b} \right)$$

As dependence among a fixed number of observations increases, the contribution of new information made by each observation decreases, and therefore effective sample size decreases as well.

3.3 Estimation of Population Variance

In data analysis settings the true value of σ^2 is unknown, and must be estimated in order to facilitate inference. Under independence, the sample variance s^2 is unbiased for σ^2, where $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. However, when are observations are dependent, s^2 is not unbiased for σ^2, and instead tends to underestimate the true variance.

$$\begin{aligned} E[s^2] &= \frac{1}{n-1} E\left[\sum_{i=1}^{n} Y_i^2 - 2\bar{Y} \sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} \bar{Y} \right] \\ &= \frac{n\sigma^2}{n-1} - \frac{n \cdot var(\bar{Y})}{n-1} \\ &= \frac{n\sigma^2}{n-1} - \frac{n}{n-1} \left(\frac{\sigma^2}{n} + \frac{b}{n} \right) \\ &= \sigma^2 - \frac{b}{n-1}. \end{aligned}$$

Since $b > 0$ for any positive correlation, $E[s^2] < \sigma^2$.

Underestimating the true variance aggravates the issue of underestimating the

standard error. The amount of bias, $B_{\sigma^2}[s^2]$, ranges from $B_{\sigma^2}[s^2] = 0$ when observations are all independent, to a magnitude of σ^2 when all observations are perfectly correlated (see Appendix for derivation). If the amount of dependence in a sample grows slower than a rate of order n, this bias will go to 0 with sample size. In all of our simulation settings this source of bias is trivial. However, in settings in which b is large compared to the sample size, using s^2 for inference results in even more highly inflated statistical significance and lower coverage probabilities than in the scenario where σ^2 is known.

3.4 Results from Simulations

We will illustrate the consequences for inference about μ of ignoring the dependence between observations through the results of two simulated network dependence scenarios. For each type of dependence, we will construct networks using the latent space model described in Section 2.1 and generate dependent outcomes using either the peer influence or the latent variable procedures described in Sections 2.2 and 2.3, respectively. We run 300 simulations for each dependence setting considered, and for each simulation, we generate $n = 100$ outcome observations, $Y_1, ..., Y_{100}$. We calculate 95% confidence intervals for the mean μ using the sample average of our observations, $\bar{Y}$, the standard error of $\bar{Y}$ estimated by s^2 under the assumption of independence, and the assumption that $\bar{Y}$ is approximately normally distributed. Since we have simulated these outcomes to be identically distributed from a $N(0, 1)$ distribution, and because μ is not affected by dependence (see Section 2.3), we know that the population mean is equal to zero. We will estimate the coverage of 95% confidence intervals for each setting as the proportion of the 300 simulations that yield 95% confidence intervals covering zero.

In order to highlight the problem of ignoring the covariance terms included in b

when estimating the standard error of $\bar{Y}$, we assume for now that σ^2 is known. Incorrectly estimating σ^2 with the sample variance, s^2, would only exacerbate the problems we demonstrate below since s^2 always underestimates σ^2 when positive dependence is present in observations sampled from a network.

3.4.1 Results from latent variable dependence setting

We ran 300 simulations under each dependence level considered in this setting: $\rho = 0.0$ (corresponding to the setting where sample observations are truly indpendent), $0.25, 0.55$, and 0.85. For each simulation, we generate a network population of 200 subjects, where each subject i is assigned a value of some latent covariate, X_i, and an outcome, Y_i; for every $i \in [1, ..., 200]$

$$(X_i, Y_i) \overset{i.i.d.}{\sim} BVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

We then take a snowball sample of $n = 100$ subjects, with our initial observation sampled from the most highly connected subject (the node with the maximum degree). For each 100-observation snowball sample, we construct a 95% confidence interval for μ as if the outcome observations are independent by

$$\bar{Y} \pm \Phi_{0.975} \frac{\sigma}{\sqrt{n}}.$$

We then estimate the coverage probabilities for each the four settings separately, as the proportion of the 300 95% confidence intervals that cover 0, the true value of μ. As the correlation (ρ) increases between the latent space variable and the outcome, subjects who share a tie tend to have more highly correlated outcomes. Since we construct a snowball sample by only adding new subjects into the sample if they have

an immediate relationship to someone already in the sample, we expect outcomes gathered via a snowball sample to be more highly correlated for higher values of ρ.

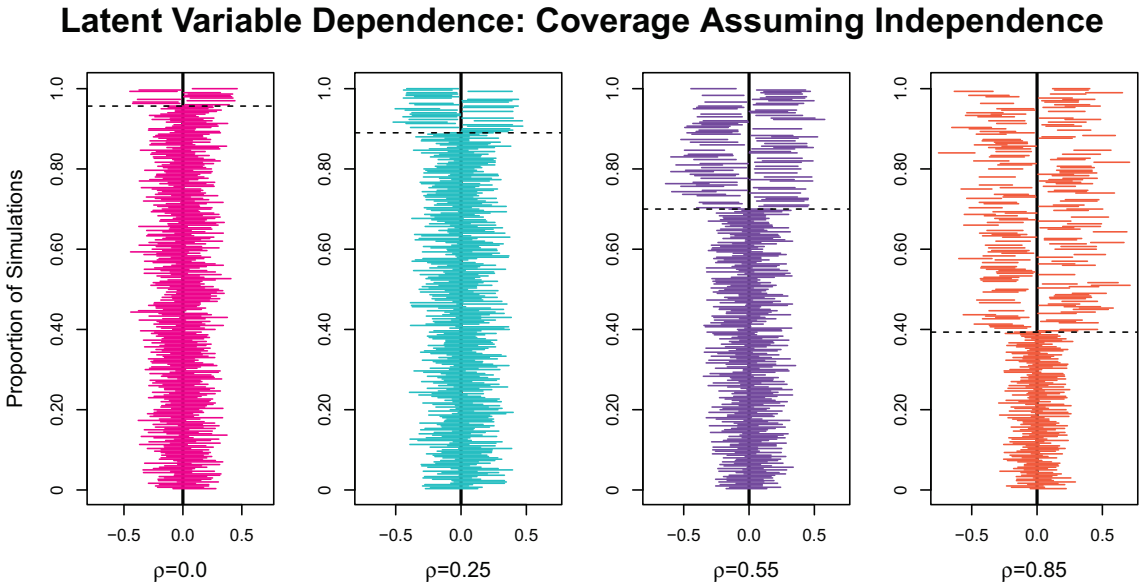

Figure 1: This figure plots 95% confidence intervals constructed from each latent variable dependence simulation assuming that observations are independent, separately by dependence level ($\rho = 0.0, 0.25, 0.55$, and 0.85). The confidence interval widths are measured along the x-axis (each horizonal line represents a confidence interval) and the y-axis measures the proportion of confidence intervals. The solid, bold, black vertical line at $x = 0$ represents the true mean, μ. The confidence intervals are sorted such that those that contain μ are plotted first (at the bottom), and those that do not contain μ are plotted after (at the top). A dashed horizonal black line is drawn where this change (from those intervals that cover μ to those that do not) occurs. This line represents our estimated coverage probability. We can see that coverage decreases substantially as correlation increases.

Unsurprisingly, we found that as dependence in a sample becomes large, the coverage probability decreases dramatically. Under independence, when $\rho = 0$, coverage was very close to 95%, at 95.66%. When ρ increased to 0.25, the coverage dropped slightly, to 89.00%. Coverage fell to 70.00%, with $\rho = 0.55$, and finally dropped to a bleak 39.33% when $\rho = 0.85$. Figure 1 illustrates these declining coverage probabilities by plotting the confidence intervals. We can see that as dependence increases, the

confidence interval widths fail to grow commensurately with the increasing variance in our sample estimate, $\bar{Y}$, and therefore fail to attain 95% coverage for μ.

We can make equivalent statements about the increasing type I error rate as ρ increases. Our simulation results indicate that we will incorrectly reject the null hypothesis that $\mu = 0$ close to the nominal rate of 5% of the time when the observations are independent, but type I error increases to 11% when $\rho = 0.25$, to 30% when $\rho = 0.55$, and to 60.66% when $\rho = 0.85$. Strange as it may sound, when dependence becomes large in this setting, we would do better to flip a coin to perform this hypothesis test rather than to statistically test for significance assuming independence.

3.4.2 Results from Peer Influence Scenario

In the peer influence setting, subjects' outcomes are initally generated to be independent observations from a $N(0, 1)$ distribution. We simulate a sequence of "interactions" among subjects over time; at each round t of interaction, the outcome for each subject i becomes a weighted average of his previous outcome and the average of each of his neighbors' previous outcomes (see Section 2.2 for details). For this setting, we generate only one underlying network structure (ensuring that network density is held constant across simulations), but assign new independent outcomes from a $N(0, 1)$ distribution at time 0 for each simulation, and then allow the subjects to interact for $t = 0, 30, 60$, and 90 rounds. We ran 300 simulations for each of four values of t, and in each simulation we calculated a 95% confidence interval for μ, incorrectly assuming that observations are independent.

We estimated coverage probabilities for each of the four settings as the proportion of the 300 95% confidence intervals that covered the true value of μ, namely 0. Under

independence, at time 0, we were able to attain 96.66% coverage, which we expected would be around 95%. At time 30, coverage drops significantly to 56.00%. At time 60, coverage decreases to 39.33%, and drops again to 33.00% at time 90. We plotted the confidence intervals (Figure 2) to demonstrate this astounding drop in coverage. The plot also illustrates how narrow the intervals become in comparison to the true variation in $\bar{Y}$ when there is dependence present between observations which is not accounted for. We underestimate the standard error of $\bar{Y}$, and therefore the widths of the confidence intervals do not grow commensurately with the variation in $\bar{Y}$ as dependence in $Y_1, .., Y_n$ increases

By the same token, the type I error rate (which should be bounded at 5%), is sufficiently low at time 0, at 3.3%. However, type I error spikes to 44% at time 30, increases again to 60% at time 60, and reaches a deplorable type I error rate of 67% at time 90. This imples that at time 90, for instance, we expect that we will incorrectly reject the null hypothesis that $\mu = 0$ 67% of the time. Hypothesis testing with highly dependent data is effectively worthless when the dependence is unaccounted for, as we make the wrong decision more often than not, and p-values carry very little value.

We present the results from these simulations because they demonstrate the detrimental impact of making statistical inferences from dependent data without accounting for the dependence. Failing to account for dependence leads us to underestimate the standard error of $\bar{Y}$ which results in low coverage and high type I error. We saw that when dependence is high ($\rho = 0.85$ in the latent space dependence scenario, or $t = 90$ in the peer influence setting), decisions made by standard hypothesis testing are completely discredited because they lead us to the wrong decision more often than the right. Coverage probabilities decrease dramatically, and confidence intervals in

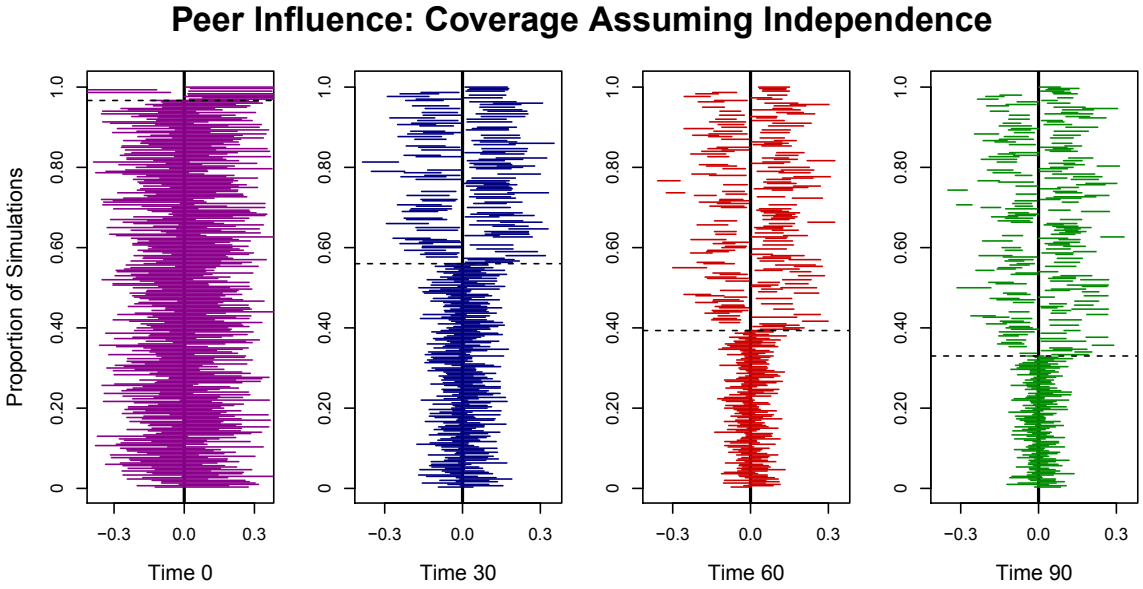

Figure 2: This figure plots 95% confidence intervals constructed from each peer influence simulation assuming that observations are independent, separately by dependence level (times $t = 0$, 30, 60, and 90). The confidence interval widths are measured along the x-axis (each horizonal line represents a confidence interval) and the y-axis measures the proportion of confidence intervals. The solid, bold, black vertical line at $x = 0$ represents the true mean, μ. The confidence intervals are sorted such that those that contain μ are plotted first (at the bottom), and those that do not contain μ are plotted after (at the top). A dashed horizonal black line is drawn where this change (from those intervals that cover μ to those that do not) occurs. This line represents our estimated coverage probability. We can see that coverage decreases substantially as correlation increases. Notice that when dependence is high, yet we fail to account for it, the width of the intervals become very small relative to the variance of our estimate $\bar{y}$ resulting in low coverage.

these settings only contain the true population parameter about a third of the time. This is potentially dangerous, because clinical and other pivotal decisions may be made from incorrectly analyzing dependent network data, and we have just shown that the inferences made are far from statistically sound. Although these points are obvious from a theoretical standpoint, they seem to go unrecognized in much of the applied literature that makes use of network observations for statistical inference.

Now that we have demonstrated the repercussions of incorrectly treating network observations as independent, we shift our focus to developing possible solutions for this issue.

4 Methods for Accounting for Dependence

We have demonstrated that it is problematic to draw inferences from network data without accounting for dependence. In this section, we will propose methods for quantifying and accounting for dependence in two different data analysis settings: first, when we have access to outcome data from multiple measured variables with equivalent dependence structure, either sampled from the same network or from multiple i.i.d. networks, and second, when we only have outcome data from a single dependent sample. While the second case is generally more common, we will see that the quantities needed to correctly estimate the standard error of estimators for μ (defined in sections below) tend to be much more stable when we have access to multiple data sets. For both cases, we will discuss methods for accounting for dependence, and then demonstrate their performance in simulations.

Recall from Section 3.1 that the major difference between inference for a population mean using independent versus dependent data is how we calculate the variance

of $\bar{Y}$ (and therefore its standard error). When we have a sample of $Y_1, ..., Y_n$ dependent observations,

$$var(\bar{Y}) = \frac{\sigma^2 + b}{n},$$

and when we incorrectly assume the observations are independent, we will underestimate $var(\bar{Y})$ by a factor of $\left(1 + \frac{b}{\sigma^2}\right)$. We denote this underestimation factor by F and refer back to this quantity in later sections. Estimating b, is not a trivial task; even in the most straightforward case when we have constructed networks and simulated outcome data ourselves, the value of b is unknown.

We first will demonstrate that we can accurately estimate b using outcomes from a very large (300) collection of independent networks. We provide this evidence solely as a proof of concept that valid inference is possible, even from very highly dependent data. We will then provide methods for inference using a small number of sets (2-5) of observations with equivalent dependence structures. This kind of data may be more feasible to collect in practice, whereas data from 300 independent networks would often be impossible to collect. We demonstrate that with only two data sources, we can approximately estimate b, resulting in valid, or almost valid, inference. The methods that we present for inference about μ using at least two data sources generally perform better than the method we propose below for the research setting in which the only available data is on one set of observations (namely, for the outcome of interest) from a single network; however, we will demonstrate that we can still improve coverage of μ by attempting to estimate b.

4.1 Case I: Inference using multiple data sets

In this section, we explore inference about the population mean of an outcome, Y, from network observations in the case when we have collected multiple (M) sets of observations to facilitate inference. Our first and foremost goal in Section 4.1.1 is to demonstrate the possibility of estimating b, thereby recovering valid inference, even for highly dependent data. By estimating b with outcomes from 300 networks, we are able to recover almost exactly 95% coverage and .05 type I error rate. While this is not necessarily representative of typical research settings, because data on many similar networks is often impossible or expensive to collect, it allows us to demonstrate that b can be well-estimated and therefore that valid inference from dependent network data is possible.

We then address a more feasible research setting, in Section 4.1.2, namely the setting in which data is available from $M = 2, .., 5$ sets of observations. In most cases, the M sets will represent outcome observations from M independent networks, under the condition that the outcome is approximately identically distributed in each network. However, we will also explore the setting where the M sets of observations correspond to measurements taken on M separate variables in a single network (with one set being the outcome of interest); in this setting we assume that the M variables are independent and have similar correlation structures. (We expect that the assumption of independence can be relaxed somewhat in this case, and we will explore this conjecture further in future work.) The estimation procedures for the standard error of $\hat{\mu}$ differ slightly for the two cases: in the former, we generally will want to use all $n \cdot M$ observations to estimate μ, b, and σ^2, while in the latter we only use all $n \cdot M$ observations to estimate b (exploiting the identical correlation structure between the M variables) and use only the n observations that correspond to the outcome of interest to estimate μ and σ^2.

4.1.1 Demonstrating the feasibility of valid inference (estimation using 300 networks)

Inference for μ requires estimates of μ, b, and σ^2. Since our primary objective in this section is to demonstrate that accurate estimation of b is possible with multiple networks and that accurate estimation of b enables us to recover valid inference about μ, we will only use data from all M independent networks to calculate b, and use data from a single network at a time to estimate μ and σ^2, and to calculate a 95% confidence interval.

Recall from Section 3.1 that we define $b = \frac{1}{n}\sum_{i\neq j}^{n} cov(Y_i, Y_j)$. Let $\underset{\sim}{Y}$ be the column vector of variables $[Y_1, Y_2, ..., Y_n]^T$. When we observe M networks, we will have M observations on each variable Y_i. We let Y_i^m be the outcome corresponding to variable $i \in [1, ..., n]$, in the m^{th} network ($m \in [1, .., M]$), and $\bar{Y_i}$ be the average value of variable i across all M networks. For this collection of M networks, we compute a single estimate of b by

$$\hat{b} = \frac{1}{n}\sum_{i\neq j}[\widehat{cov}(\underset{\sim}{Y})]_{i,j},$$

where $cov(\underset{\sim}{Y})$ is the variance-covariance matrix of $\underset{\sim}{Y}$. That is, $\widehat{cov}(\underset{\sim}{Y})$ is the matrix with entry (i, j) equal to $\frac{1}{M-1}\sum_{m=1}^{M}(Y_i^m - \bar{Y_i})(Y_j^m - \bar{Y_j})$, $i, j = 1, ..., n$. We present results from simulations below demonstrating that $\hat{b}$ is a good estimate of b in many settings, and even for small M (in Section 4.1.2). When data is available from multiple similar networks we would usually expect node identities and network configuration to be different across networks, and this makes it difficult or impossible to meaningfully or stably estimate the variance-covariance matrix for $\underset{\sim}{Y}$- although we have M repeated

observations of $\underset{\sim}{Y}$, there is no meaningful sense in which Y_i^m from one sample is equivalent to Y_i^k in another, and therefore no stable interpretation of the off-diagonal entries in the estimated matrix $\widehat{cov}(\underset{\sim}{Y})$. However, because b is a sum over all of the entries in the variance-covariance matrix, we can estimate b when the underlying structure of each network is entirely different (e.g. each network is a new draw from the data-generating process), not just when the network structures are identical. That is, it is not necessary for any subject i in network k to have the same set of relationships as subject i in network m in order to accurately estimate b. This is because $\hat{b}$ simplifies to an expression devoid of any operation over subject i across networks. We can express $\hat{b}$ as follows:

$$\hat{b} = \frac{1}{n}\sum_{i\neq j}\left[\frac{1}{M}\sum_{m=1}^{M} Y_i^m Y_j^m - \frac{1}{M^2}\sum_{k=1}^{M} Y_i^k \sum_{k=1}^{M} Y_j^k\right] \tag{1}$$

$$= \frac{1}{n}\frac{1}{M}\sum_{m=1}^{M}\sum_{i\neq j} Y_i^m Y_j^m - \frac{1}{n}\frac{1}{M^2}\sum_{m=1}^{M}\sum_{k=1}^{M}\sum_{i\neq j} Y_j^m Y_i^k. \tag{2}$$

The innermost operations in expression (2) above fix the network identities at m and k and sum over all pairs $(i, j : i \neq j)$. Only after all individual level data has been aggregated do we sum over networks. This indicates that $\hat{b}$ has no structure; it is invariant to permutations of the identities of the nodes in any or every network. Therefore, even though we may not have the type of data that would permit robust and stable estimation of a highly structured variance-covariance matrix, we are able to estimate the unstructured summary measure b.

The sample mean, $\bar{Y}$ is an unbiased estimate for $\mu = 0$, regardless of the correlation among outcomes (see Section 2.3). To estimate μ in the m^{th} network, we will use $\bar{Y}_m$, the sample mean of observations in network m. We showed in Section 3.3 that $\sigma^2 = E[s^2] + \frac{b}{n-1}$ under dependence. For each network, m, we estimate $E[s^2]$ with s_m^2 (s^2 computed only from observations in network m), and an estimate for $\hat{b}$. For

network, m,

$$\hat{\sigma_\rho^2} = s_m^2 + \frac{\hat{b_\rho}}{n-1}.$$

We calculate 95% confidence intervals for each network or simulation, m, separately by

$$\bar{Y}_m \pm \Phi_{0.975}\sqrt{\frac{\hat{\sigma_\rho}^2 + \hat{b_\rho}}{n}},$$

or equivalently,

$$\bar{Y}_m \pm \Phi_{0.975}\sqrt{\frac{s_m^2}{n} + \frac{\hat{b_\rho}}{n-1}}.$$

We estimate coverage probabilities by the proportion of the M confidence intervals that cover $\mu = 0$.

Below we demonstrate the performance of this procedure using data from simulated networks with dependence due to latent variable and peer influence. For every dependence level, we will estimate b, calculate F (the factor by which we underestimate the variance of $\bar{Y}$ when incorrectly treating the observations as independent, defined earlier in Section 4), and estimate the coverage probability of the resulting 95% confidence intervals. (See Appendix for evidence that the methods for accounting for dependence are robust to initially non-normally distributed, continuous outcomes.) Additionally, we will estimate average effective sample size for each dependence setting, and use those estimates to generate quantile-quantile plots providing evidence that our assumption from Section 3, that $\bar{Y}$ converges to a normal distribution, was reasonable (even when observations are heavily dependent).

We begin with the latent variable dependence simulation setting in which the outcome of interest, Y, is correlated with the latent variable, X, according to which edges form in the network, and dependence in each observed network is realized through a snowball sample of $n = 100$ subjects from a total population of 200. For each subject, $i = 1...100$,

$$(X_i, Y_i) \overset{i.i.d.}{\sim} BVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

For each simulation, we generated an overall network of 200 subjects, and collected a snowball sample of $n = 100$ observations. We performed 300 such simulations for each of four separate dependence settings: $\rho = 0.0, 0.25, 0.55$, and 0.85. When $\rho = 0$, we know that $b = 0$ and $\sigma^2 = 1$; however, for $\rho > 0$, we do not know the true values of b and σ^2, but we expect that the estimates provided here, calculated using observations across 300 independent networks are very close to their true values. Since correlation is positive in the latter three scenarios, we expect that $\hat{b}_{.25}$, $\hat{b}_{.55}$, and $\hat{b}_{.85}$ will all be greater than zero (where $\hat{b}_\rho$ represents the estimate of b in the setting where the correlation between the latent variable X and the outcome Y is defined to be ρ), and we estimated

$$\hat{b}_{0.0} = .02$$
$$\hat{b}_{.25} = 0.38$$
$$\hat{b}_{.55} = 1.88$$
$$\hat{b}_{.85} = 4.87.$$

We expect $\hat{\sigma^2_\rho}$ to decrease with increasing dependence in the network: as dependence increases, observations become more similar, and therefore the marginal variance of the outcome decreases. For each network, m, under dependence setting defined by ρ,

we estimate σ_ρ^2 with $\hat{\sigma_\rho^2} = s_m^2 + \frac{b_\rho}{n-1}$, and below we report the estimates averaged over 300 simulations each:

$$avg(\hat{\sigma_0^2}) = 0.99$$
$$avg(\hat{\sigma_{.25}^2}) = 0.95$$
$$avg(\hat{\sigma_{.55}^2}) = 0.77$$
$$avg(\hat{\sigma_{.85}^2}) = 0.46.$$

In this setting, the bias of s^2 for σ^2 was fairly trivial. Figure 3 is a plot of the distribution of s^2 across the 300 simulations for the four values of ρ. The center of mass of these distributions is approximately equal to $E[s_\rho^2]$, and visual inspection shows that each curve is approximately centered around the corresponding mean value of $\hat{\sigma_\rho^2}$ listed above, implying that the bias of s^2 for σ^2, in this case, is not very large. We plot vertical dashed lines in Figure 3 representing $avg(\hat{\sigma}_\rho^2)$ for each ρ to show the amount of bias of s^2 for σ^2.

Under the assumption that all observations are independent, we would conclude that $var(\bar{Y})_\rho = \frac{\sigma_\rho^2}{n}$, and we would underestimate the true $var(\bar{Y})_\rho$ by a factor of $F = \left(1 + \frac{b_\rho}{\sigma_\rho^2}\right)$. Table 1 presents average estimates for the variance of $\bar{Y}$, both assuming observations are independent and accounting for the dependence, and the average value of F across 300 simulations. As ρ increases, the estimates of $var(\bar{Y})$ under the assumption of independence are strictly decreasing, but estimates accounting for dependence are strictly increasing. Therefore, the factor by which estimates failing to account for dependence are off (F) increases, and the consequences of incorrectly assuming observations are independent become more grave: the width of the confidence interval for μ shrinks by a factor of $\sqrt{\left(1 + \frac{b_\rho}{\sigma_\rho^2}\right)}$ relative to the valid 95% confidence interval, which inevitably decreases the probability that it will cover the truth. We

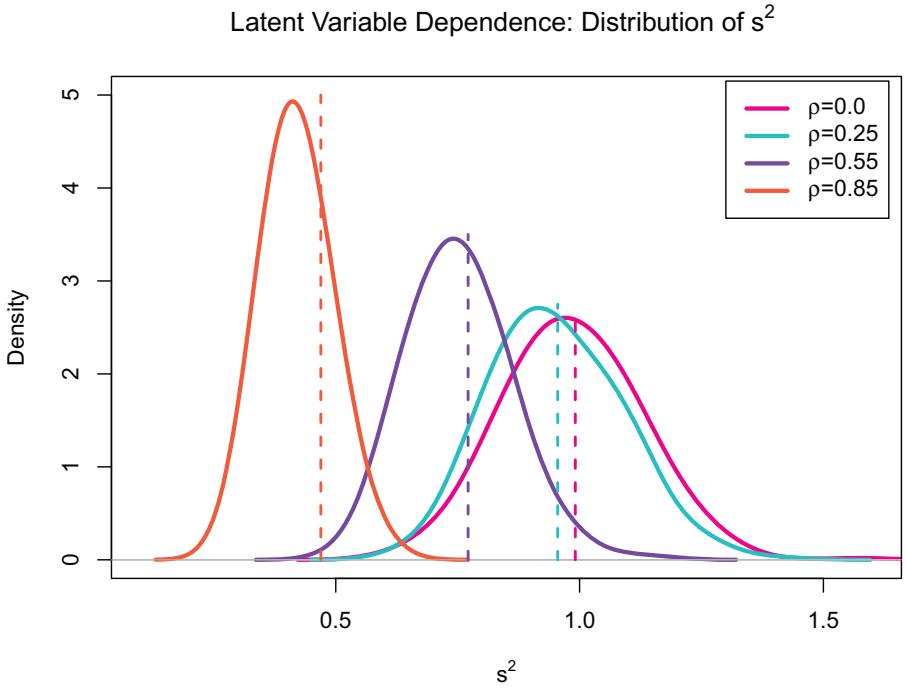

Figure 3: This figure shows the distributions of s_m^2 for a snowball sample of 100 observations when $\rho = 0$ (in pink), when $\rho = 0.25$ (in blue), when $\rho = 0.55$ (in purple), and when $\rho = 0.85$ (in orange). We see that $E[s_\rho^2]$–the center of mass of the distributions–decrease with increasing correlation. We found that the stimated $E[s_\rho^2]$ is approximately equal to the average of estimates for σ_ρ^2, which are plotted in the corresponding colors in dashed vertical lines on top of the density plots. This indicates that in this setting, the bias of s_m^2 for σ_ρ^2 is rather trivial, especially when dependence is small.

presented the consequences of failing to account for this dependence in section 3.4. When we account for dependence using the estimates of b given above, we can achieve 95% coverage, even with very high dependence present in the sample. We computed 95% confidence intervals for μ for each single network, m,

$$\bar{Y}_m \pm \Phi_{0.975} \frac{\hat{\sigma_\rho}}{\sqrt{n}} \cdot \sqrt{\left(1 + \frac{b_\rho}{\hat{\sigma_\rho^2}}\right)}.$$

These adjusted confidence intervals achieve 94.66% coverage or higher for all dependence settings (95.66% when $\rho = 0$, 95.66% when $\rho = 0.25$, 94.66% when $\rho = 0.55$,

and 95.66% when $\rho = 0.85$). This result is illustrated in Figure 4, where we plot the corrected confidence intervals; we see that approximately 95% of the intervals cover $\mu = 0$ regardless of the amount of dependence (compare to Figure 1, above). Equivalently, type I error is bounded at $\alpha \approx 0.05$. While this implies that we can still draw accurate inferences from dependent observations, note that inference is not as precise as it would be with independent data drawn from the same marginal distribution. When observations are correlated, our confidence intervals will have to widen proportionally with the amount of the dependence in order to ensure sufficient coverage.

Variance of $\bar{Y}$ Accounting For Vs. Ignoring Dependence

Correlation	Average $var(\bar{Y})$ Assuming Independence	Average $var(\bar{Y})$ Accounting for Dependence	Average Factor $=\left(1 + \frac{b_\rho}{\sigma^2_\rho}\right)$
0.0	0.0099	0.0102	1.02
0.25	0.0097	0.0135	1.40
0.55	0.0087	0.0276	3.49
0.85	0.0068	0.0555	11.56

Table 1: This table presents the estimated variance of $\bar{Y}$ first assuming observations are independent, and then when accounting for the dependence for four dependence settings by estimating b and incorporating the estimate in the computation of $\text{var}(\bar{Y})$. The last column presents the average factor, F, across the 300 simulations for each dependence setting ($\rho = 0.0, 0.25, 0.55$, and 0.85).

For each dependence setting, that is for $\rho = 0.0, 0.25, 0.55$, and 0.85, and for a sample of $n = 100$ observations, we can estimate the effective number of independent observations that would lead us to draw the exact same inference about μ, or the effective sample size n_e. Recall from Section 3.2 that $n_e(\rho) = n \cdot \left(\frac{\sigma^2_\rho}{\sigma^2_\rho + b_\rho}\right)$. We

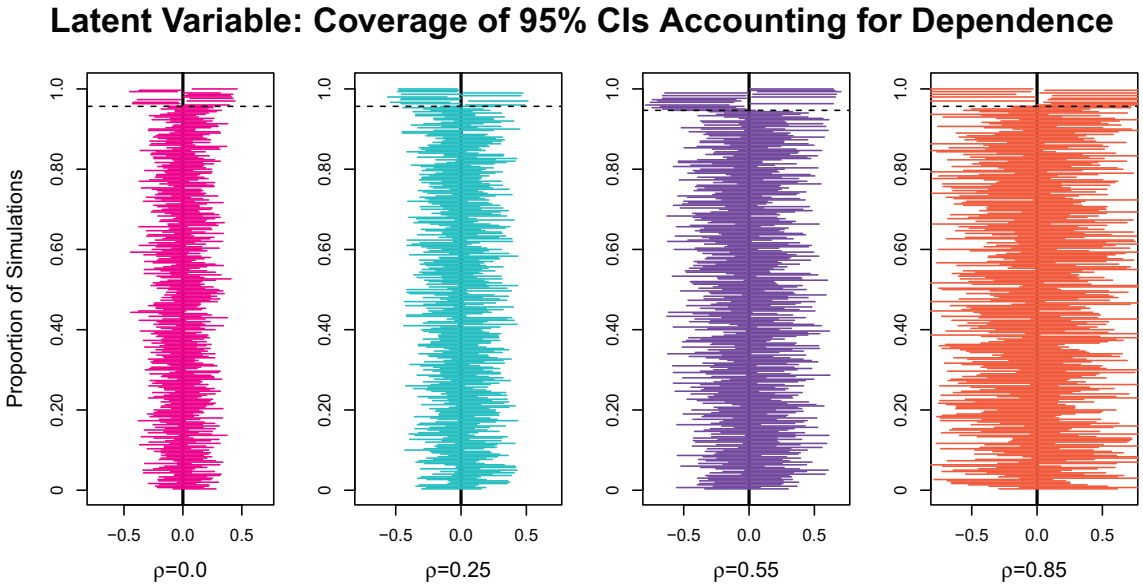

Figure 4: Here we have 95% confidence intervals (drawn horizontally and measured along the x-axis) which take the dependence between observations into account, separately for dependence level. We see that for all levels of dependence, coverage (marked by the horizontal dashed black line) hovers right around 95%. Notice that in order to ensure this high coverage, however, the width of the intervals increase as dependence becomes larger.

compute the average estimated effective sample size for each value of ρ:

$$avg(\hat{n}_e(0.0)) = 97.37$$

$$avg(\hat{n}_e(0.25)) = 71.14$$

$$avg(\hat{n}_e(0.55)) = 28.89$$

$$avg(\hat{n}_e(0.85)) = 8.77$$

These effective sample sizes demonstrate the inefficiency of dependent data relative to independent observations; a sample of 100 observations when $\rho = 0.85$ is equivalent to a random sample from a population with the outcome having the same marginal outcome distribution less than a tenth of our original sample size. Under the assumption that using our dependent data, $\bar{Y} \sim N\left(0, \frac{\sigma_\rho^2 + b_\rho}{n}\right)$ (the assumption is that $\bar{Y}$ is approximately normally distributed when $n = 100$), we should be able to draw the same inferences from independent observations, $Z_1, Z_2, ...Z_{n_e(\rho)}$, where $Z_i \sim N(0, \sigma_\rho^2)$ and therefore $\bar{Z} \sim N\left(0, \frac{\sigma_\rho^2}{n_e(\rho)}\right)$. We made quantile-quantile plots comparing these two distributions (of observed $\bar{Y}$ and generated $\bar{Z}$) for each dependence setting to demonstrate their equivalence (Figure 5). The quantile-quantile plots not only demonstrate that $n_e(\rho)$ is the correct effective sample size for independent observations, but they show that $\bar{Y}$ is, in fact, approximately normally distributed since the distributions are approximately equivalent.

We now shift our focus to the results produced under the peer influence setting. In the peer influence setting, we generate a single underlying latent space network structure, and for each simulation we generate dependence in the sample by allowing the subjects to "interact" for a given number of discrete time points. The outcome for subject i at any time t, is given by a weighted average of his outcome and the average of his neighbor's outcome values at time $t-1$. We report results from 300 simulations

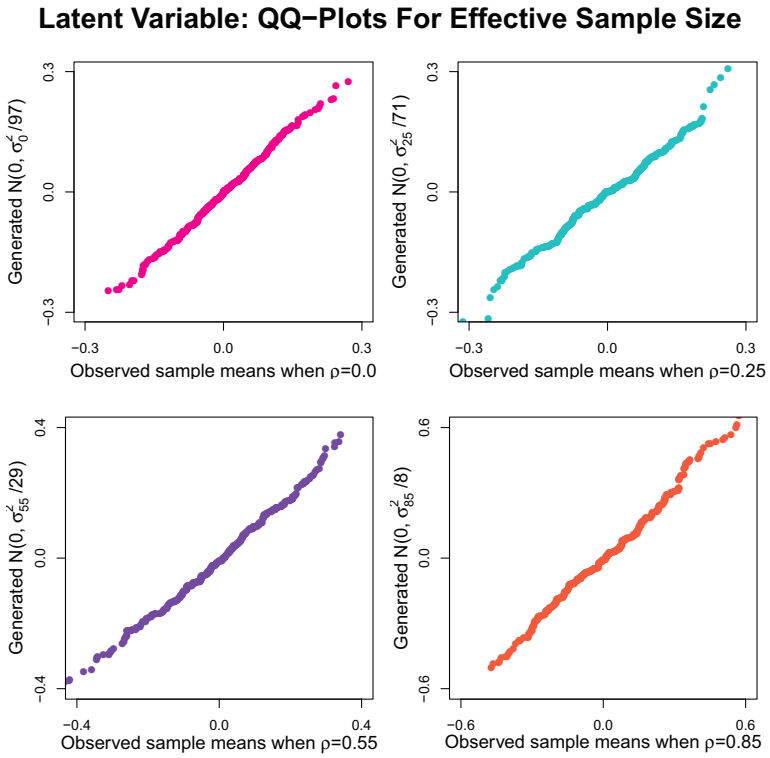

Figure 5: Here we have four quantile-quantile plots which compare the distribution of the observed sample means computed from 100 dependent observations (x-axis) with the distribution of generated random variables representing the sample mean of $n_e(\rho)$ independent observations, from a population with the same marginal mean and variance, σ^2_ρ (y-axis). We can see that the quantiles match up in each plot, which shows that the two distributions being compared in each quarter of the figure are approximately equivalent; this provides evidence that a central limit theorem holds for our dependent data in this setting.

for each of four values of t ($t = 0, 30, 60$, and 90). We account for dependence by computing a single estimate of b for each value of t, pooling information from the 300 networks simulated under that dependence setting.

We expect $b \approx 0$ at time 0., when all observations are independent However, since correlation is positive for any time $t > 0$, we expect that $\hat{b}_{30}$, $\hat{b}_{60}$, and $\hat{b}_{90}$ will all be greater than zero (where $\hat{b}_t$ represents the estimate of b in the setting where the subjects in a network have interacted for t rounds), and we estimated

$$\hat{b}_0 = 0.00$$

$$\hat{b}_{30} = 0.82$$

$$\hat{b}_{60} = 0.95$$

$$\hat{b}_{90} = 1.01$$

If we let the subjects interact over the course of enough rounds, all outcomes would converge to a common value. Figure 6 demonstrates that we are beginning to see some convergence of the outcome values around zero at time 90 (note that this is just an example of how the distribution of Y changes over time within *one* simulated network). In general, we expect that $\sigma_0^2 > \sigma_{30}^2 > \sigma_{60}^2 > \sigma_{90}^2$ (where σ_t^2 represents the variance of Y at time t). We estimate σ_t^2 separately for each simulated network sample, m, by $\hat{\sigma_t^2} = s_m^2 + \frac{\hat{b}_t}{n-1}$. We report estimates for σ_t^2, averaging over the 300 simulations ran under each dependence setting below:

$$avg(\hat{\sigma_0^2}) = 1.00$$

$$avg(\hat{\sigma_{30}^2}) = 0.15$$

$$avg(\hat{\sigma_{60}^2}) = 0.07$$

$$avg(\hat{\sigma_{90}^2}) = 0.05$$

As above, in this setting the bias of s_m^2 for σ_t^2 was again negligible. Figure 7 plots the distribution of s_m^2 across the 300 simulations for the four values of t. The center of mass of these distributions is approximately equal to $E[s_t^2]$, but each curve is also approximately centered around the corresponding average value of $\hat{\sigma_t^2}$ listed above, indicating that the bias of s_m^2 for σ_t^2 is relatively small under these settings. We plot vertical dashed lines in Figure 7 representing $avg(\hat{\sigma}_t^2)$ for each t considered to show

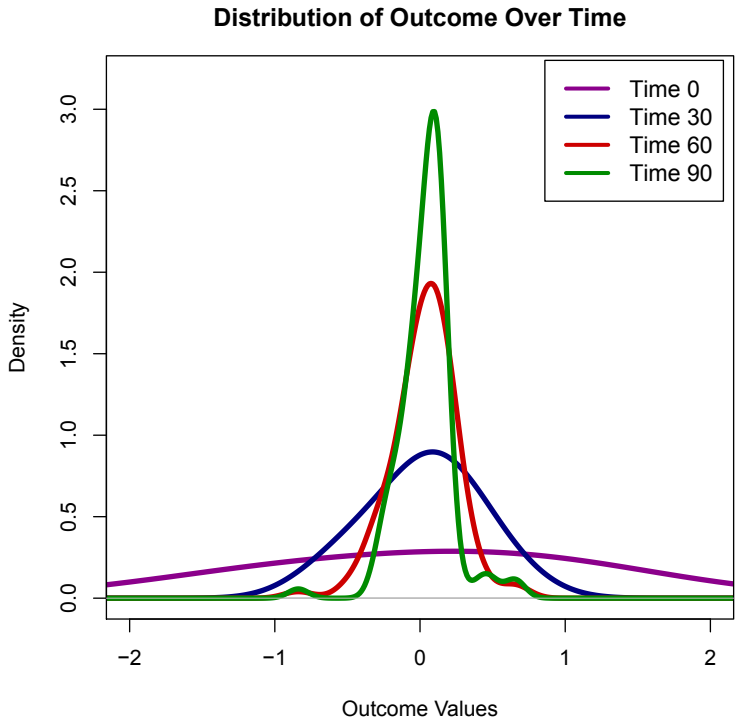

Figure 6: This figure shows how the distribution of our outcome, Y, changes over time. We've plotted the distribution of Y when all obervations were independent at time 0 (in purple), at time 30 (in blue), at time 60 (in red) and at time 90 (in green). Initially $Y_i \sim^{iid} N(0,1)$, but we can see that over time, observations become more and more similar, and seem to be converging on a common value around zero (note that it is not always the case that they will converge around zero). Therefore, we'd expect that the variance of Y at each time point becomes smaller than it was previously.

the amount of bias of s^2 for σ^2.

Since positive correlation is present in the sample after time 0, failing to account for this dependence would lead us to underestimate $var(\bar{Y})$ by a factor of $F = \left(1 + \frac{b_t}{\sigma_t^2}\right)$. Table 2 shows similar trends to the results under latent variable dependence reported above: as t increases, we observe decreasing estimates for $var(\bar{Y})$ when we incorrectly assume observations are independent, increasing estimates for $var(\bar{Y})$ when correctly accounting for dependence using estimates for b, and therefore increasing factors, F by which we underestimate $var(\bar{Y})$ when failing to account for dependence.

Variance of $\bar{Y}$ Accounting For Vs. Ignoring Dependence			
Time Point	Average $var(\bar{Y})$ Assuming Independence	Average $var(\bar{Y})$ Accounting for Dependence	Average Factor $=\left(1+\frac{b_t}{\hat{\sigma}_t^2}\right)$
0	0.0100	0.0100	1.00
30	0.0016	0.0098	6.36
60	0.0008	0.0103	15.02
90	0.0006	0.0107	20.71

Table 2: This table presents the average variance of $\bar{Y}$ first assuming observations are independent, and then when accounting for the dependence for time 0, time 30, time 60, and time 90. We see that as correlation increases, the $\bar{Y}$ values which assume observations are independent are strictly decreasing, but the $\bar{Y}$ values which account for dependence tend to increase. Therefore, the factor F (the ratio of $\bar{Y}$ accounting for dependence to $\bar{Y}$ assuming independence) grows with increasing correlation, resulting in more severe consequences for inference.

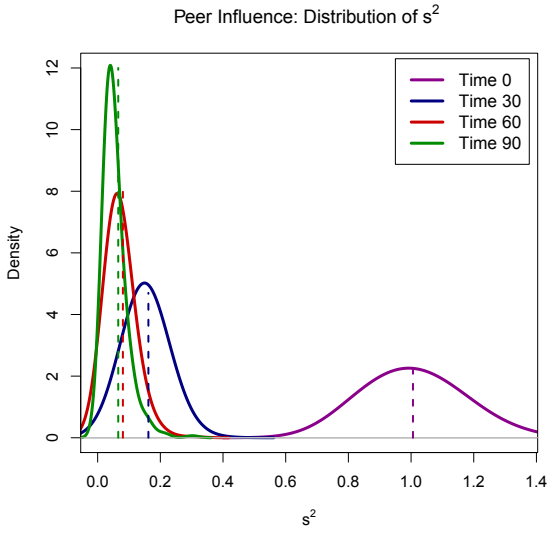

Figure 7: This figure shows the distributions of s_m^2 for a network of 100 observations when subjects have been interacting for a varying number of time points in the peer influence setting. We plot s_m^2 when $t = 0$ (in purple), when $t = 30$ (in blue), when $t = 60$ (in red), and when $t = 90$ (in green). We see that $E[s_t^2]$–the center of mass of the distributions–decreases with increasing time points. We found that the estimated $E[s_t^2]$ is approximately equal to the average of estimates for σ_t^2, which are plotted as dashed vertical lines in the corresponding colors on top of the density plots. This indicates that in this setting, the bias of s_m^2 for σ_t^2 is rather trivial, especially when dependence is small.

The factors by which we underestimate $var(\bar{Y})$ when we incorrectly assume independence between observations determine the severity of the consequences of failing to account for dependence for inference about μ. When we fail to correct for the correlation present in the sample, the confidence intervals that we construct for μ shrink by a factor of $\sqrt{\left(1 + \frac{b_t}{\sigma_t^2}\right)}$ relative to the valid 95% confidence interval, which decreases the probability that it will cover the truth. That is, the confidence intervals produced are inappropriately narrow for the large variation in $\bar{Y}$ when the sample contains dependence, resulting in low coverage, as we demonstrated in Section 3.4. When we account for dependence using the estimates of b_t given above, we can attain approximately 95% coverage, even with very high dependence present in the sample. We computed 95% confidence intervals for μ for each single network, m,

$$\bar{Y}_m \pm \Phi_{0.975} \frac{\hat{\sigma}_t}{\sqrt{n}} \cdot \sqrt{\left(1 + \frac{\hat{b}_t}{\hat{\sigma_t^2}}\right)}.$$

Using our estimates of b to account for dependence in this construction of confidence intervals, we achieved 94% coverage or higher for all dependence settings: we estimated 96.66% coverage at time 0, 95.00% coverage at time 30, and 94% coverage at both time 60 and time 90. The confidence intervals widen proportionally with the increased variation in $\bar{Y}$ when obserations in samples are correlated to ensure sufficient coverage; we plot the confidence intervals, accounting for dependence, in Figure 8 to demonstrate this phenomenon (compare to Figure 2) .

We can equate our 100 dependent observations to some n_e independent observations with the same marginal distribution, by the degree of inferential precision we have. Recall that we can estimate $n_e(t)$ by $\hat{n}_e(t) = n \cdot \left(\frac{\hat{\sigma}_t^2}{\hat{\sigma}_t^2 + \hat{b}_t}\right)$. We estimate $n_e(t)$ for each network sample under the four dependence settings $t = 0, 30, 60$, and 90; below, we list the estimated effective sample size averaged across the 300 simulations ran

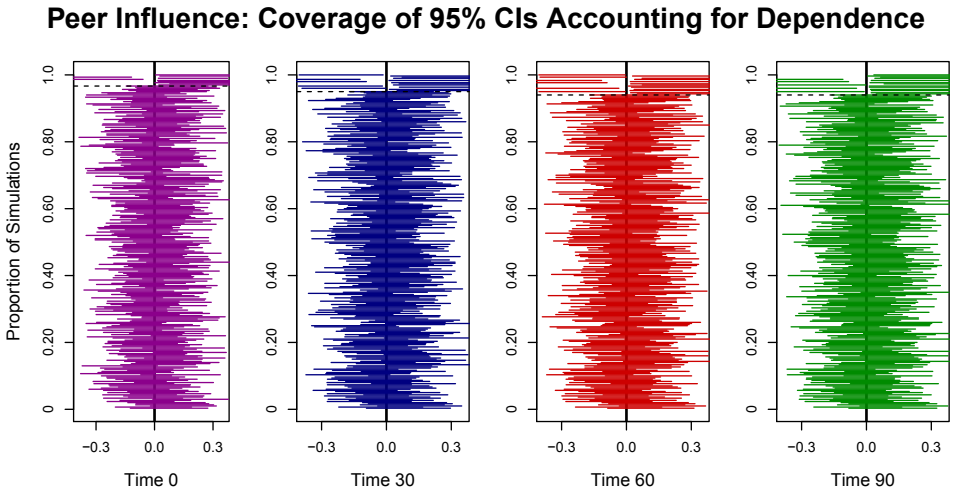

Figure 8: This figure shows 95% confidence intervals for the four time points examined, when taking dependence into account. We see that we are able to achieve approximately 95% coverage for all dependence settings (denoted by the blacked dashed horizontal line). Notice that the intervals are much wider here (except at time 0) than the were in Figure 2, which showed the confidence intervals assuming that observations were independent.

under each dependence setting:

$$avg(\hat{n_e}(0)) = 100.00$$

$$avg(\hat{n_e}(30)) = 15.95$$

$$avg(\hat{n_e}(60)) = 7.17$$

$$avg(\hat{n_e}(90)) = 5.49$$

These estimates for effective sample size are dramatically lower than our dependent sample size ($n = 100$), demonstrating the inefficiency of dependent data compared to a sample of independent observations from the same marginal distribution. We see that a sample of (less than) six independent observations from a population where the outcomes, $X_1, ..., X_6 \overset{i.i.d}{\sim} N(0, \sigma^2_{90})$ would produce the same inference for μ as our dependent sample of 100 observations at time 90. Under the assumption that using our dependent sample of $n = 100$, $\bar{Y} \sim N\left(0, \frac{\sigma^2_t + b_t}{n}\right)$, we should be able to draw

the same inferences about μ from independent observations, $X_1, X_2, ...X_{n_e(t)}$, where $X_i \sim N(0, \sigma_t^2)$ and therefore $\bar{X} \sim N\left(0, \frac{\sigma_t^2}{n_e(t)}\right)$. We made quantile-quantile plots comparing these two distributions of observed $\bar{Y}$ and randomly generated $\bar{X}$ for each dependence setting (t=0, 30, 60, and 90) to demonstrate that they are approximately equivalent, even with high dependence. This not only validates our estimates for effective sample size, but they show that $\bar{Y}$ is approximately normally distributed. We present the results from these simulations to demonstrate that it is possible to

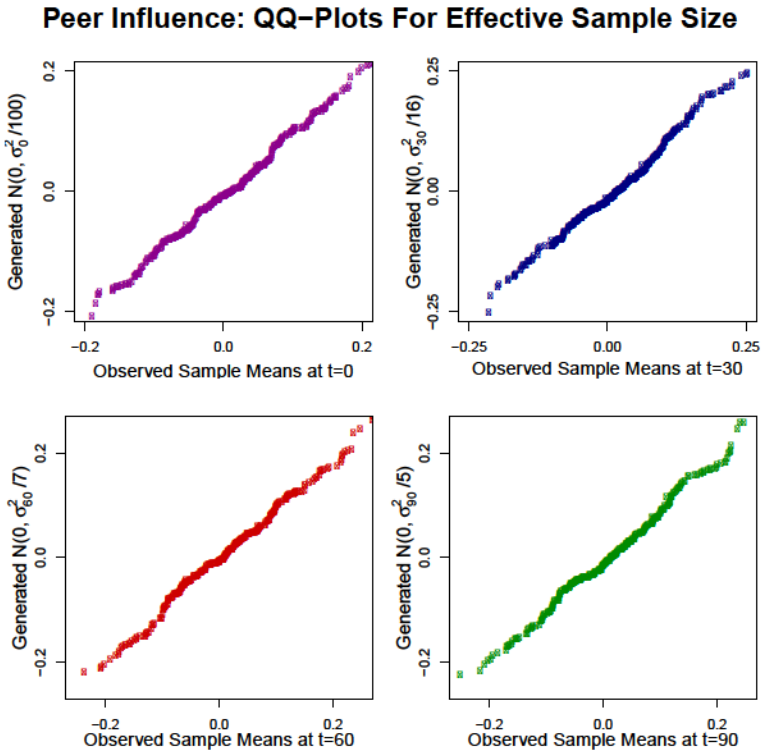

Figure 9: Here we have four quantile-quantile plots which compare the distribution of observed sample means formed from 100 dependent observations (x-axis) with the distribution of generated random variables representing the sample means of $n_e(t)$ independent observations, from a population with the same marginal distribution (y-axis). We can see that the quantiles match in each plot, which shows that each of the two distributions being compared are approximately equivalent, and therefore that a central limit theorem still holds for $\bar{Y}$.

accurately estimate the quantity b, and therefore draw accurate inferences from dependent network data, but only when correctly accounting for the dependence. (See Appendix for evidence that the methods for accounting for dependence are robust

to initially non-normally distributed, continuous outcomes.) While we lose some inferential precision by accounting for the dependence between observations, we gain accuracy which is usually more important (a precise confidence interval, no matter how narrow, tends to have very little benefit if it does not contain the population parameter of interest). We will now explore the more realistic scenario in which data is available on only $M = 2, .., 5$ networks or variables, rather than on 300 independent networks.

4.1.2 Inference Using Few Networks

While we demonstrated that accurate inference for the population mean, μ, of an outcome of interest, Y, is possible when we have access to outcome data from 300 independent networks, this is not feasible or practical in most research settings. Here we implement the methods proposed above on simulated data from a small number of independent networks or independent outcomes collected from the same network ($M = 2, .., 5$), and we believe that all variables/networks have approximately the same correlation structures. We ran $300 \cdot M$ simulations for each dependence setting (times $t = 0, 30, 60, 90$, and ρ=0.0, 0.25, 0.55, and 0.85) to construct 300 separate confidence intervals for each setting. We use information from M unique networks or variables in the construction of each confidence interval. We will estimate coverage by the proportion of confidence intervals which cover the truth, and will demonstrate that we can achieve valid, or almost valid, inference using only a relatively small number of samples.

When we have sampled outcome observations from multiple networks, and believe that the outcome is approximately identically distributed in each network, then it is sensible to use all of the information gathered to produce the most accurate confidence

interval. Here, M represents the number of independent networks from which we have sampled outcome observations and n represents number of subjects in each network. To maximize accuracy, we incorporate all $M \cdot n$ outcome observations in the estimation methods for μ, b, and σ^2. Let $\bar{Y_m}$ be the sample mean of the outcome from network m, s_m^2 be the sample variance of the outcome from network m, and $\underset{\sim}{Y}$ be the column vector of variables $[Y_1, Y_2, ..., Y_n]^T$ (where for M networks sampled, we will have M observations on each variable Y_i). Then

$$\hat{\mu} = \frac{1}{M} \sum_{m=1}^{M} \bar{Y_m} = \bar{\bar{Y}}$$

$$\widehat{E[s^2]} = \frac{1}{M} \sum_{m=1}^{M} s_m^2 = \bar{s^2}$$

$$\hat{b} = \frac{1}{n} \sum_{i \neq j} [\widehat{cov}(\underset{\sim}{Y})]_{i,j},$$

The greater the number of networks used to estimate b, the more stable the estimate will be. Due to the very small number of independent networks we use to estimate b here, the estimate is relatively unstable and it's possible that it would sometimes lead us to incorrectly conclude that there is negative correlation in the sample ($b < 0$). Because we are only considering social networks with positive correlation, we make the following assumption: either the network observations are positively correlated, or they are independent (in other words, ruling out the possibility that the observations are negatively correlated regardless of what the data shows). Therefore, if we happen to estimate b to be less than zero, we set $\hat{b}$ to zero and conclude that the observations are independent.

Using the estimates $\hat{b}$ and $\bar{s^2}$, we can estimate σ^2 by

$$\hat{\sigma^2} = \bar{s^2} + \frac{\hat{b}}{n-1}.$$

Recall from Section 3.1 that we can express the estimated variance of $\bar{Y}$ in terms of $\hat{\sigma^2}$ and $\hat{b}$ as

$$\widehat{var(\bar{Y})} = \frac{\hat{\sigma^2} + \hat{b}}{n}$$

However, in this case we are considering $\bar{\bar{Y}}$ as the point estimate for μ, and therefore we need to estimate $var(\bar{\bar{Y}})$ instead:

$$\begin{aligned} \widehat{var(\bar{\bar{Y}})} &= \frac{1}{M}\widehat{var(\bar{Y})} \qquad \text{(since networks are independent)} \\ &= \frac{\bar{s^2}}{M \cdot n} + \frac{\hat{b}}{M(n-1)} \end{aligned}$$

Therefore, when we have identically distributed outcome data from multiple independent networks, we construct our 95% confidence intervals for μ by

$$\bar{\bar{Y}} \pm \Phi_{0.975}\sqrt{\frac{\bar{s^2}}{M \cdot n} + \frac{\hat{b}}{M(n-1)}}.$$

Coverage calculated from confidence intervals constructed in this manner are listed in Tables 3 and 4 under the column "Coverages: Multiple Networks." In the context of both the peer influence and the latent variable dependence simulation settings, we found that we can achieve greater than 82% coverage for all dependence settings (times $t = 0, 30, 60, 90$, and ρ=0.0, 0.25, 0.55, and 0.85) using outcome observations from only four networks. With outcome observations sampled from five networks, we were able to achieve 84.66% coverage or higher for the peer influence setting, and 89.66% coverage or higher for the latent variable dependence scenario. Since we

previously found that ignoring the dependence inherent in these networks results in coverage probabilities as low as 33%, it would clearly be worthwhile, when possible, to collect data from a couple of additional networks in order to accurately quantify and account for dependence, as this can significantly improve coverage.

Researchers will not always have access to multiple networks, but perhaps they can collect data on various characteristics from the same network. If one of more of the additional variables measured has approximately the same correlation structure as the outcome of interest, we can scale the variable(s) such that the range matches the outcome of interest, and use the additional variable(s) to achieve a more stable estimate of b than would be possible with only the single set of outcome observations. We will show below that using this auxiliary information solely to estimate b can produce fairly accurate estimation of b, resulting to more accurate inference than is possible when the only data available is data from a single variable in a single network (see Section 4.2, below). We estimate b in the same manner as we did previously, but now M refers to the number of variables collected from a network, m indexes the different variables measured, and we will let o refer to the set of data corresponding to the outcome of interest. In this case we use the data on all variables ($M \cdot n$ total observations) to estimate b, but we only use the n observations which correspond to the outcome of interest to estimate μ or $E[s^2]$, our estimates for the following quantities become

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} Y_i^o = \bar{Y_o} \\ \widehat{E[s^2]} &= s_o^2 \\ \hat{b} &= \frac{1}{n}\sum_{i \neq j} [\widehat{cov}(\underset{\sim}{Y})]_{i,j}. \end{aligned}$$

We maintain the assumption that observations are either positively correlated or

independent, and will set $\hat{b} = 0$ if this method yields a negative estimate for b. The final column of Tables 3 and 4 show how many estimates (out of 300) were incorrectly estimated to be negative and manually set to zero. Using $\hat{b}$ and s_o^2, our estimate for σ^2 is

$$\hat{\sigma^2} = s_o^2 + \frac{\hat{b}}{n-1}.$$

Then, solving for the variance of $\bar{Y_o}$,

$$\begin{aligned}\widehat{var(\bar{Y_o})} &= \frac{\hat{\sigma^2} + \hat{b}}{n} \\ &= \frac{s_o^2}{n} + \frac{\hat{b}}{n-1}.\end{aligned}$$

Therefore, when we have collected data on multiple variables that we believe have the same correlation structure as our outcome of interest, we construct our 95% confidence intervals for μ by

$$\bar{Y_o} \pm \Phi_{0.975}\sqrt{\frac{s_o^2}{n} + \frac{\hat{b}}{n-1}}.$$

Coverage estimated from confidence intervals constructed in this manner are listed in Tables 3 and 4 under the column "Coverages: Multiple Variables." Recall that even though the circumstances were framed in a different way, this is exactly how we constructed confidence intervals using 300 simulations (the purpose before was only to demonstrate that b could be accurately estimated using a large number of independent networks). For the peer influence simulation setting we were able to attain 90% coverage or higher for all dependence settings (times t=0, 30, 60, and 90) using observations from only three variables, and 91% or higher with five variables. For the latent variable dependence setting, we were able to achieve 92% coverage or higher for all dependence settings ($\rho = 0.00, 0.25, 0.55, 0.85$) using four networks, and 95.66%

coverage or higher with five networks. Coverage is much improved compared to the procedures that fail to account for dependence (which resulted in coverage as low as 33%), and we will eventually see (in Section 4.2) that it is also an improvement over procedures that use only a single network to estimate b.

Although the second method in which M represents variables from a single network generally results in higher coverage, there are some advantages to using all $M \cdot n$ observations to estimate μ and $E[s^2]$, in addition to b. The point estimate for μ in the first method described (for M representing multiple networks), $\bar{\bar{Y}}$, tends to be much more accurate than only using the sample mean from one set of data, since we average across multiple independent networks. Additionally, the standard error for $\bar{\bar{Y}}$ is generally much smaller than the standard error of one sample mean, resulting in much tighter confidence intervals and more precise inference. This is evidenced by the mean squared error (MSE) for our estimator, $\hat{\mu}$, which we estimate by

$$\widehat{E[(\hat{\mu} - \mu)^2]} = \widehat{var(\hat{\mu})} + \hat{\mu}^2.$$

For every collection of M networks under each dependence setting in the case when M represents multiple variables from a single network, we compute an estimate of MSE, and denote these estimates by MSE_{var}; that is, for every dependence setting (t=0, 30, 60, 90 and ρ=0.0, 0.25, 0.55, and 0.85) we calculate 300 estimates of MSE_{var}. Similarly, for every collection of M networks under each dependence setting in the case when M represents multiple independent networks from which we have sampled outcome observations, we compute an estimate of MSE, and denote these estimates by MSE_{net}; that is, for each of the eight dependence settings we calculate 300 estimates of MSE_{net}. We compare the MSE estimates for the two methods by calculating the average ratio of MSE_{var} to MSE_{net} for each dependence setting. We

present this average ratio in Tables 3 and 4 under the column heading "Avg. $\frac{MSE_{var}}{MSE_{net}}$." For both the peer influence and the latent variable dependence settings, we found that the method for multiple variables results in an MSE approximately 2 times larger than the method for multiple networks with two data sets, approximately 3 times larger for 3 data sets, approximately 4 times larger for 4 data sets, and 5-6 times larger for 5 data sets. While using observations from multiple networks to estimate all quantities for inference (μ, $E[s^2]$, and b) results in slightly lower coverage in comparison to the multiple variable method (only using multiple data sets to estimate b), the benefit is much more precise inference.

Due to the assumption that we previously made–that observations are either positively correlated or independent–when we try to account for (non-existant) dependence by estimating b from only a few networks in which observations are truly independent, such as at time 0 or when $\rho = 0.0$, we will tend to get conservative inference. This is due to the fact that, although we expect our b estimates to be centered around zero, the estimates will randomly deviate from zero in either direction (the magnitude determined in part by the number of networks used). Negative estimates will always be set to zero by our procedure, but all positive estimates (indicating positive correlation present) remain. This results in unnecessarily widened confidence intervals and therefore conservative coverage for independent observations.

Peer Influence: Coverages Using Multiple Data Sets					
Sets	Time	Coverage: Networks	Coverage: Variables	Avg. $\frac{MSE_{var}}{MSE_{net}}$	# $\hat{b} < 0$
2	0	96.66	99.00	2.12	200
	30	79.33	87.00	2.15	81
	60	76.66	82.33	2.14	47
	90	73.00	78.00	2.12	46
3	0	95.00	99.00	3.30	191
	30	83.33	91.33	3.42	37
	60	81.66	91.00	3.38	17
	90	81.00	90.33	3.36	13
4	0	98.00	99.33	3.65	194
	30	90.33	95.66	4.81	25
	60	89.66	93.66	4.81	11
	90	82.66	89.33	4.90	25
5	0	97.00	99.33	5.70	188
	30	89.33	94.33	5.79	15
	60	89.33	93.66	5.75	5
	90	84.66	91.00	6.39	15

Table 3: Coverages calculated using outcome data from 2-5 independent networks, or 2-5 variables from a single network, the MSE ratio comparing the two methods, and the number of negative b estimates under the peer influence scenario.

Latent variable dependence: Coverages Using Multiple Data Sets					
Sets	ρ	Coverage: Networks	Coverage: Variables	Avg. $\frac{MSE_{var}}{MSE_{net}}$	# $\hat{b} < 0$
2	0.0	98.00	99.66	2.06	196
	0.25	92.66	97.33	2.06	184
	0.55	79.00	88.33	2.04	113
	0.85	75.33	82.00	2.08	69
3	0.0	97.33	99.33	3.35	189
	0.25	93.00	96.33	3.30	141
	0.55	87.00	91.33	3.38	67
	0.85	79.66	89.00	3.27	24
4	0.0	96.33	99.00	4.51	196
	0.25	94.33	98.33	4.34	143
	0.55	86.33	93.00	4.66	34
	0.85	82.66	92.00	4.45	9
5	0.0	97.33	99.00	5.81	196
	0.25	92.33	95.66	5.56	122
	0.55	90.33	97.33	5.89	24
	0.85	89.33	95.66	5.73	2

Table 4: Coverages calculated using outcome data from 2-5 independent networks, or 2-5 variables from a single network, the MSE ratio comparing the two methods, and the number of negative b estimates under the latent variable dependence scenario.

Now that we have demonstrated the performance of the proposed methods for accounting for dependence when we have collected multiple sets of relevant observations that allow us to learn about b, we will address the more challenging case: drawing inferences about μ when we have sampled outcome observations from only one dependent network.

4.2 Case 2: Inference from a single dependent network sample

In this section, we describe inference about μ using n outcome observations from a single network, $Y_1, ..., Y_n$. Estimating b with only one sample is much more challenging than the settings we discussed above. We will take an alternative approach to estimating b; unlike $\hat{b}$ using multiple sets of observations (described in Section 4.1), this method takes into account the structure of the sampled network. As above, we assume that observations are either positively correlated or independent. Although we tend to substantially underestimate b in single sample estimates, which results in anti-conservative coverage, we will demonstrate that inference is far more accurate and coverage is much improved compared to inference that ignores dependence altogether.

4.2.1 Estimating quantities for correction

Recall that $b = \frac{1}{n}\sum_{i \neq j} cov(Y_i, Y_j)$ is the scaled sum of all of the pairwise covariance terms for Y. In settings where we observe multiple networks, or multiple independent outcomes on the same network, we are able to use information from the independent samples of the dependence structure to obtain stable estimates of this quantity. In contrast, a single sample $Y_1, ..., Y_n$ often contains little and unreliable information about the underlying covariance structure. One of the primary reasons for this is that we have to center our estimates of $cov(Y_i, Y_j)$ around the observed sample mean

$\bar{Y}$, but $\bar{Y}$ is a highly variable estimate of μ, especially under dependence. Furthermore, whenever the sample distribution is symmetric we expect approximately equal numbers of pairs of observations to lie on the same side of $\bar{Y}$ and to lie on opposite sides of $\bar{Y}$, that is we expect approximately half of the pairwise covariance terms to be negative–even if, in truth, every pair is positively correlated. In order to circumvent these issues, we abandon the task of estimating b and instead attempt the less ambitious goal of learning about the approximate magnitude of b.

In social networks, we expect that the majority of dependence is attributable to the correlation between outcomes from immediate friends and that as the distance between a pair of nodes increases, observations from that pair contribute a decreasing amount to the total sum of pairwise covariance terms. This need not always be the case, but it is a reasonable assumption for any kind of dependence that is informed by network ties, and it holds for the peer influence and latent variable dependence-generating mechanisms that we consider in our simulations. To learn about the magnitude of the measure of overall dependence, b, we will calculate separate measures of dependence for each distance between subjects possible in the network, and then sum the measures only for those subject distances that give positive measures of dependence to obtain a very rough estimate of b. We estimate the covariance between observations from any two subjects i and j by determining how similarly the two observations deviate from the sample mean. When there is positive dependence in a network, we expect, for instance, that if one subject has an outcome value much larger than the network average, then his close friends will also have high outcomes relative to the sample mean. Note that we can express b as a sum of terms corresponding to

subject distance

$$\hat{b} = \frac{1}{n}\left[\sum_{i,j:dist(i,j)=1} \widehat{cov}(Y_i, Y_j) + ... + \sum_{i,j:dist(i,j)=k} \widehat{cov}(Y_i, Y_j)\right]$$
$$= \frac{1}{n}\left[\sum_{i,j:dist(i,j)=1} (Y_i - \bar{Y})(Y_j - \bar{Y}) + ... + \sum_{i,j:dist(i,j)=k} (Y_i - \bar{Y})(Y_j - \bar{Y})\right]$$

where $dist(i, j)$ represents the length of the shortest path from i to j, and we choose k to be the largest degree of separation between friends such that the sum of estimated covariance terms is positive. k can range from $k = 0$, in which case we fail to detect any positive correlation in the network, to the largest observed pairwise distance in the network (the "diameter" of the network), in which case we estimate that all pairs of subjects, separated by any distance, are positively correlated. Although it is often realistic for all pairs of observations to be positively correlated, and in fact holds for both of our dependence-generating simulation settings, for the reasons cited above it is generally impossible to detect using observed data. Instead, we observe that terms are decreasing in distance, and terms for distance greater than some small k are negative. This is simply a consequence centering the covariance estimates around the sample mean and does not necessarily imply that subjects separated by $> k$ degrees are negatively correlated. Because the terms are decreasing in k, to omit all negative terms from the sum it suffices to sum over k only up to the first distance that results in a negative term.

To better understand the limitations of this estimation method, consider the simple example where we gather data from a dense network of 15 dependent observations. The outcome values for all 15 subjects and network structure for this example are illustrated in figure 10. We can see from the figure that the maximum amount of separation between any two subjects is two edges ($max(dist(i, j)) = 2$), and as-

sume we know that all observations (between all pairs of subjects, $dist(i,j) = 1$ and $dist(i,j) = 2$) are positively correlated. Since subjects separated by only one edge (where $dist(i,j) = 1$) are more highly correlated than subjects separated by two edges (where $dist(i,j) = 2$), immediate friends are more likely to deviate in the same direction. Consequently, the outcomes of pairs of subjects where $dist(i,j) = 2$ must lie on opposite sides of the sample mean; by our estimation method, we would incorrectly conclude that subjects of distance 2 are negatively correlated (when, in fact, they are *positively* correlated), and exclude these terms from our estimate of b (in other words, we would set $k = 1$ when in reality pairs with $k = 2$ are positively correlated as well). As a consequence of our estimation method, we only account for the dependence attributable to friends of distance k or less, when in reality friends of distance $> k$ could very well be positively correlated; this tends to lead to an underestimate of b.

The amount by which we underestimate b is at least in part affected by the density of the network. Consider a very dense network of subjects, where all subjects have immediate relationships to each other (all are first-degree friends); in this case, even when observations are truly very positively correlated, we are just as likely to estimate that the observations are independent or negatively correlated. In this case, we implement the assumption that observations are either independent or *positively* correlated; therefore, if we happen to estimate that $b < 0$, we will set $b = 0$, and assume observations are independent. Because we occasionally incorrectly assume that observations are independent, and tend to underestimate b when $b > 0$, this method generally leads to anti-conservative standard error estimates. However, accounting for dependence in this manner still greatly increases coverage probabilities in all dependence settings explored, in comparison with assuming independence in all cases.

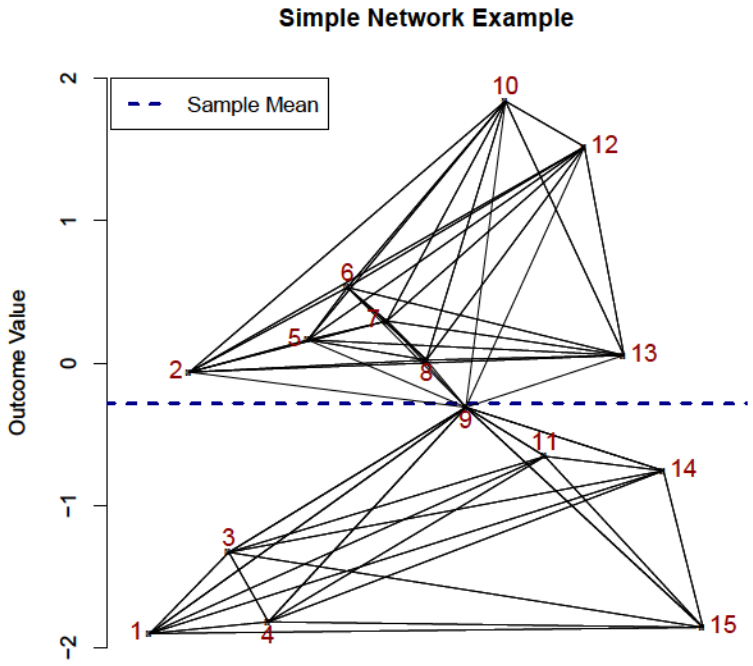

Figure 10: This figure shows a simple network of 15 subjects illustrating the limitations of our estimation method for b. The dark red numbers represent unique subject IDs, and the black lines represent ties between friends; as we can see, all subjects are separated by either one or two degrees. All subjects (both first and second degree friends) are positively correlated. However, in dense networks such as this, we incorrectly estimate that second-degree friends are negatively correlated, and only include the pairwise covariance terms for first-degree friends in our estimation for b.

Given an estimate for b, we then calculate an estimate of σ^2 for the single network of interest, by

$$\hat{\sigma}^2 = s^2 + \frac{\hat{b}}{n-1}.$$

where s^2 represents the sample variance of the outcome. We then construct 95%

confidence intervals for any single network by

$$\bar{Y} \pm \Phi_{0.975}\sqrt{\frac{\hat{\sigma^2} + \hat{b}}{n}}.$$

We now demonstrate that we can achieve improved coverage using this method compared to assuming observations are independent, by implementing this method for accounting for dependence in simulations, beginning with the latent variable dependence setting.

For each of the four values of ρ (0, 0.25, 0.55, 0.85) considered in the latent variable dependence setting described in 2.3, we simulated 300 latent space networks of 200 subjects, and obtained snowball samples of $n = 100$ outcomes from each. We estimated single sample estimates for b as described above, and constructed a 95% confidence interval for μ for each snowball sample. Even though we expected that single sample estimates of b would lead to anti-conservative inferences, we still found that coverage probabilities, estimated as the proportion of the 300 confidence intervals that covered 0, improve greatly compared to the coverage of 95% confidence intervals constructed incorrectly assuming independence. When we account for dependence with single sample estimates, we found that coverage increased from 89.00% to 95% when $\rho = 0.25$, from 70.00% to 82% when $\rho = 0.55$,and from 39.33% to 82.33% when $\rho = 0.85$. Under independence we saw no notable change in coverage (we estimated 96% coverage when $\rho = 0$). The confidence intervals from all simulations are plotted in Figure 11 to illustrate the increase in the width of the intervals relative to those constructed assuming independence, which are depicted in figure 1.

Although there are some obvious limitations of this method that do not allow us to always achieve 95% coverage, our inferences are vastly improved compared to the

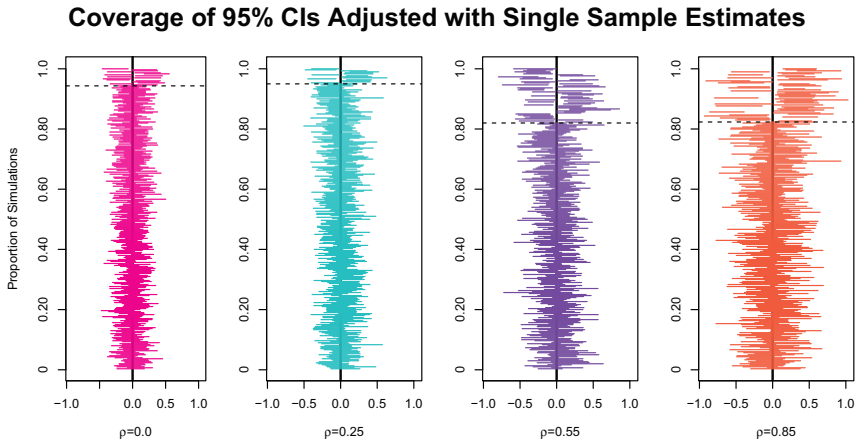

Figure 11: This figure plots 95% confidence intervals for 300 simulations under each dependence setting: $\rho = 0$ (observations are independent), $\rho = 0.25$, $\rho = 0.55$, and $\rho = 0.85$. The standard error for $\bar{Y}$ accounts for the dependence between observations using single sample estimates of b.

case when we always assume independence between observations. Our type I error decreases from $\alpha = 0.11$ to $\alpha = 0.05$ when $\rho = 0.25$,$\alpha = 0.30$ to $\alpha = 0.18$ when $\rho = 0.55$, and $\alpha = 0.60$ to $\alpha = 0.1766$ when $\rho = 0.85$. Accounting for dependence using single sample estimates of b has clear advantages for inference, despite the limitations of this method.

In the peer influence setting, we generate dependence over a series of time points. For each of the four time points considered (t=0, 30, 60, and 90), we generate 300 peer influence samples (as described in Section 2.2), calculate single sample estimates for b (as described above), and construct 95% confidence intervals for μ using the single sample estimate for b, $\hat{\mu} = \bar{Y}$, and $\widehat{E[s^2]} = s^2$. Due to the limitations of the single sample estimation procedure for b, we expected inference to be anti-conservative when sampling dependent observations from only one network. Although slightly anti-conservative, we found that coverage probabilities, estimated by the proportion of these 300 95% confidence intervals covering $\mu = 0$, drastically improved from coverage estimated by confidence intervals constructed assuming independence. Under

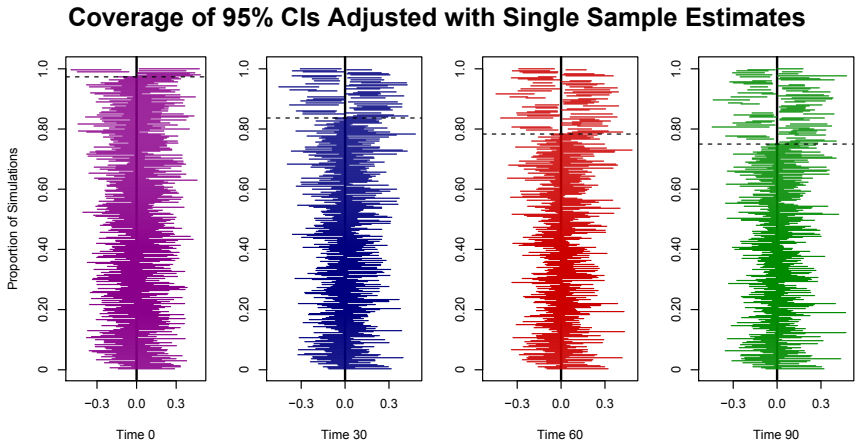

Figure 12: This figure plots 95% confidence intervals for 300 simulations under each dependence setting: time 0 (observations are independent), time 30, time 60, and time 90. The standard error for $\bar{Y}$ accounts for the dependence between observations using single sample estimates of b.

independence, at $t = 0$, coverage remained sufficiently high at 97%. Coverage improved from 56% to 83.66% at $t = 30$, from 39.33% to 78.33% at $t = 60$, and from 33.00% to 75% at t =90. As expected, we didn't quite achieve 95% coverage, but the benefits of using single sample estimates to account for dependence are clear and significant from the immense improvements in coverage. We plot the 95% confidence intervals in figure 12 to demonstrate that coverage improves due to improved estimation of standard error, resulting in wider confidence intervals in comparison to the confidence intervals constructed assuming independence, illustrated previously in Figure 2.

We have seen that our method for estimating b using only a single network still tends to underestimate the standard error of $\bar{Y}$ for most dependence settings, and therefore produces confidence intervals with lower coverage and higher type I error than desired. However, we have shown from our simulation results that this method provides obvious benefits for inference over incorrectly assuming independence between observations. While data from multiple networks is ideal for precisely quan-

tifying and accounting for dependence in inference for μ, we have found that this method provides a reasonable alternative for inference when outcomes can only be sampled from a single network.

5 Future Work

Moving forward, we plan to address the limitations of the estimation methods that we propose here for b, both when employing outcome observations from multiple networks (or multiple variables sampled from the same network) and when only sampling dependent outcome observations from a single network. When we have sampled outcome observations from M independent networks (or M independent variables with similar correlation structures from the same network), an immediate limitation of the expression we suggest for $\hat{b}$ (defined in Section 4.1) is the requirement that all M networks (or variables) must have equal sample size, n. When M represents observations on multiple variables sampled from the same network, this condition will not pose a problem. However, when M represents multiple independent networks from which we have sampled outcome observations, there is no guarantee that the networks will be similar in size, let alone exactly equal. We would like to either modify the estimation method for b to allow for differing sample size by network, or to explore methods for, and the implications of, reducing sample size in all networks to the smallest network sampled to implement the current estimation method. When we observe M variables from the same network, we plan to evaluate the impact of relaxing the requirement of independence between variables; we expect that our estimation procedure will be robust to slight dependence. When sampling dependent outcome observations a single network, the most significant limitation of the method we propose for computing $\hat{b}$ is its tendency to underestimate b. Underestimating b causes us to underestimate the standard error of $\bar{Y}$, leading to anti-conservative inference. Potential modifications

to overcome this problem are an important target of future research.

It will be important to generalize these methods to other types of data and alternative research objectives. The research objective that we addressed in this work is statistical inference about a population mean from a sample of dependent network data where, prior to generating dependence, each subject's outcome was normally distributed. We developed inferential methods (for μ) to account for dependence and demonstrated that these methods greatly improve the accuracy of inference with our specific data and we provide some preliminary evidence (in Appendix) that these methods are robust to inferring information about μ when the initial outcomes (prior to generating dependence) are non-normal, but still independent and continuous. We would like to continue exploring the performance of these methods on outcomes following alternative continuous distributions, perhaps those having skewed or otherwise unfavorable qualities. We will address these future objectives through simulations, but we plan to supplement our findings with a better understanding of their theoretical foundation; in particular, we would like to prove (if true) that a central limit theorem holds for observations correlated due to underlying network structure under certain conditions. A logical next step is to consider modifying these methods to allow for compatibility with binary outcome observations, with the goal of estimating the population prevalence of the outcome. We would also like to consider extending the methods that we suggest for accounting for network dependence to apply to inference using other population estimators (besides $\bar{Y}$ for inference about μ), such as structural characteristics or transmission behaviors in networks. As discussed in Section 1.1, network dependence is likely present in studies across a wide range of research areas; standard methods for inference in these areas should be updated to quantify and account for possible correlation between observations to allow for valid inference.

Finally, we plan to explore dependence generating patterns in alternative underlying network structures. In particular, we are interested in assessing how these results and methods extend to network samples drawn from *directed* networks, under both latent variable dependence setting and under peer influence. Directed networks bring many complications; for example it is possible for $dist(i,j) \neq dist(j,i)$ and $T_{ij} \neq T_{ji}$. Directed edges would also allow for multiple strongly connected sub-networks to form, meaning that multiple subgroups of subjects develop which cannot be affected (via peer influence) or reached (via snowball sample) by any other node outside of the corresponding subgroup even when the network comprises a single connected component when directionality is disregarded. The density of the network is largely affected by directionality; a directed network with no reciprocated edges becomes twice as dense by changing the structure to undirected. As a result, we expect that directed latent space networks, holding all else constant, might decrease the amount of dependence generated in the network. Without the proper attention to possible dependence, valid inference is often not possible.

References

[1] D. Heckathorn. "Respondent-driven sampling: A new approach to the study of hidden populations." *Social Problems*, 44: 174-199, 1997.

[2] E. Volz, D. Heckathorn. "Probability Based Estimation Theory for Respondent Driven Sampling." *Journal of Official Statistics*, 24(1): 79-97, 2008.

[3] S. Goel, M. Salganik. "Assessing respondent-driven sampling." *Proceedings of the National Academy of Sciences*, 107: 6743-6747, 2010.

[4] K. Gile, M. Handcock. "Respondent-Driven Sampling: An Assessment of Current Methodology." *Sociological Methodology*, 40(1): 285-327, 2012.

[5] K. Gile. "Improved inference for respondent-driven sampling data with application to HIV prevalence estimation." *Journal of the American Statistical Association*, 106: 135-146, 2011.

[6] S. Nesterko, J. Blitzstein. "Bias-Variance and Breadth-Depth Tradeoffs in Respondent-Driven Sampling." Submitted 22 Oct 2012.

[7] S. Lunagomez, E. Airoldi "Bayesian Inference from Non-Ignorable Network Sampling Designs." Submitted 19 Jan 2014.

[8] R. Hanneman, M. Riddle. *Introduction to Social Network Methods.* (Riverside, CA: University of California, Riverside. 2005).

[9] E. Airoldi, S. Fienberg, A. Goldenberg, A. Zheng. "A survey of statistical network models." *Foundations and Trends in Machine Learning*, 2(2):1-117, 2009.

[10] N. Christakis, J. Fowler. "The Spread of Obesity in a Large Social Network over 32 Years." *The New England Journal of Medicine*, 357: 370-379, 2007.

[11] J. Fowler, N. Christakis. "Dynamic spread of happiness in a large social network: longitudinal analysis of the Framingham Heart Study social network." *British Medical Journal*, 338: 23-27, 2009.

[12] N. Christakis, J. Fowler. "The Collective Dynamics of Smoking in a Large Social Network." *The New England Journal of Medicine*, 358: 2249-2258, 2008.

[13] R. Lyons. "The Spread of Evidence-Poor Medicine via Flawed Social-Network Analysis." *Statistics, Politics, and Policy*, 2(1), 2011.

[14] C. Latkin. "Outreach in natural settings: the use of peer leaders for HIV prevention among injecting drug users' networks." *Public Health Reports*, 113: 151-159, 1998.

[15] C. Latkin, S. Sherman, A. Knowlton. "HIV prevention among drug users: Outcome of a network-oriented peer outreach intervention." *Health Psychology*, 22(4): 332-339, 2003.

[16] M. Jurgensen, I. Sandoy, C. Michelo, K. Fylkesnes. "Effects of home-based voluntary conselling and testing on HIV-related stigma: Findings from a cluster-randomized trial in Zambia." *Social Science & Medicine*, 81: 18-25, 2013.

[17] R. Jewkes, M. Nduna, J. Levin, N. Jama, K. Dunkle, et al. "Impact of Stepping Stones on incidence of HIV and HSV-2 and sexual behaviour in rural South Africa: cluster randomised controlled trial." *British Medical Journal*, 337:a506, 2008.

[18] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, et al. "Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases." *PLoS Med*, 5(3):e74, 2008.

[19] A. Klovadhal, J. Potterat, D. Woodhouse, J. Muth, Q. Muth, et al. "Social networks and infectious disease: the Colorado Springs study." *Social Science & Medicine*, 38(1):79-88, 1994.

[20] M. Ali, D. Dwyer. "Social network effects in alcohol consumption among adolescents." *Addictive Behaviors*, 35:337-342, 2010.

[21] J. Cacioppo, J. Fowler, N. Christakis. "Alone in the crowd: The structure and spread of loneliness in a large social network." *Journal of Personality and Social Psychology*, 97(6):977-991, 2009.

[22] A. Madan, S. Moturu, D. Lazer, A. Pentland. "Social sensing: obesity, unhealthy eating and exercise in face-to-face networks." *Wireless Health*, 104-110, 2010.

[23] J.N. Rosenquist, J. Murabito, J. Fowler, N. Christakis. "The spread of alcohol consumption behavior in a large social network." *Annals of Internal Medicine*, 152(7):426-433, 2010.

[24] C.R. Shalizi, A. Thomas. "Homophily and contagion are generically confounded in observational social network studies." *Sociological Methods and Research*, 40:211-239, 2011.

[25] E. Cohen-Cole, J. Fletcher. "Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic." *Journal of Health Economics*, 27(5):1382-1387, 2008.

[26] H. Noel, B. Nyhan. "The "unfriending" problem: The consequences of homophily in friendship retention for causal estimates of social influence." *Social Networks*, 33(3):211-218, 2011.

[27] A. Neaigus, S. Friendman, R. Curtis, D. Jarlais, R. Furst, et al. "The relevence of drug injectors' social and risk networks for understanding and preventing HIV infection." *Social Science & Medicine*, 38(1): 67-78, 1994.

[28] S. Friendman, A. Neaigus, B. Jose, R. Curtis, D. Jarlais. "Networks and HIV risk: an introduction to social network analysis for harm reductionists." *Internation Journal of Drug Policy*, 9(6): 461-469, 1998.

[29] P. Hoff, A. Raftery, M. Handcock. "Latent Space Approaches to Social Network Analysis." *Journal of the American Statistical Association*, 97: 1090-1098, 2002.

Appendix

Derivation of bias

We note in Section 3.3 that the amount of bias of s^2 for σ^2, $B_{\sigma^2}[s^2]$, ranges from $B_{\sigma^2}[s^2] = 0$ when observations are all independent, to $B_{\sigma^2}[s^2] = -\sigma^2$ when all observations are perfectly correlated. When all observations are independent, it is clear that s^2 is unbiased for σ^2, but we will show here why $B_{\sigma^2}[s^2] = -\sigma^2$ when all observations are perfectly correlated.

When all observations are perfectly correlated, we have that $corr(Y_i, Y_j) = 1$ and let $var(Y_i) = var(Y_j) = \sigma^2$ for all i and j, so $cov(Y_i, Y_j) = \sigma^2$. We will first solve for b in this dependence setting for any n observations:

$$\begin{aligned} b &= \frac{1}{n}\sum_{i\neq j} cov(Y_i, Y_j) \\ &= \frac{1}{n}\sum_{i\neq j} \sigma^2 \\ &= \frac{1}{n}\left[2\binom{n}{2}\sigma^2\right] \\ &= \frac{\sigma^2}{n}\left[2 \cdot \frac{n!}{2!(n-2)!}\right] \\ &= \sigma^2\left[\frac{(n-1)!}{(n-2)!}\right] \\ &= (n-1)\sigma^2. \end{aligned}$$

We found in Section3.3 that $E[s^2] = \sigma^2 - \frac{b}{n-1}$, and in this dependence setting,

$$\begin{aligned} E[s^2] &= \sigma^2 - \frac{(n-1)\sigma^2}{n-1} \\ &= 0. \end{aligned}$$

Therefore, the bias of s^2 for σ^2 when all observation are perfectly correlated is

$$\begin{aligned} B_{\sigma^2}[s^2] &= E[s^2] - \sigma^2 \\ &= -\sigma^2. \end{aligned}$$

Simulations with Uniform and Poisson outcome observations

We explore here the robustness of the methods we propose for estimating b, and its impact on inference when the outcome of interest is still continuous, but initially non-normally distributed. We consider only the peer influence setting (see Section 2.2), with the only modification being the distribution of Y assigned to the $n = 100$ subjects at time $t = 0$. Where in the original peer influence setting we let $Y_i^0 \overset{i.i.d.}{\sim} N(0,1)$ (Y_i^0 representing the outcome for subject i at time 0), we now consider two new continuous distributions: $Y_i^0 \overset{i.i.d.}{\sim} Uniform[-1,1]$ and $Y_i^0 \overset{i.i.d.}{\sim} Poisson(5)$. Dependence is generated over time, exactly as before, and we consider the same t =0, 30, 60, and 90, with 300 simulations for each t.

We calculate 95% confidence intervals for each simulation, incorrectly assuming independence between observations, exactly as we do in Section 3.4.2. We estimate coverage for each t as the proportion of the 300 95% confidence intervals covering μ ($\mu = 0$ and $\mu = 5$ for the uniform and poisson distributions, respectively). The estimated coverages for each t from both the initally $Uniform[-1,1]$ and $Poisson(5)$ outcome distributions are strikingly similar to the coverages estimated in Section 3.4.2, when the outcome was initially i.i.d. $N(0,1)$, and are presented below in Tables 6 and 5 under the second column, "Coverage assuming independence." For every t, we then estimate b using observations from all 300 networks, and calculate adjusted 95% confidence intervals exactly as we do in Section 4.1.1 for each simulation. We estimate coverage again as the proportion of the 300 95% adjusted confidence

intervals that cover μ, and present the estimates in Tables 6 and 5 under the second column, "Coverage accounting for dependence." We find that observations need not be initially normally distributed in order to recover valid inference, as we were able to attain ≈95% coverage for every level of dependence considered.

Peer Influence: Coverages with Initally Poisson(5) Outcome		
Time	Coverage assuming independence	Coverage accounting for dependence
0	95.00	94.66
30	58.00	95.33
60	45.33	95.33
90	39.33	95.66

Table 5: Coverage estimates, both incorrectly treating observations as independent and accounting for dependence, when the outcome initially follows a $Poisson(5)$ distribution.

Peer Influence: Coverages with Initially Uniform$[-1, 1]$ Outcome		
Time	Coverage assuming independence	Coverage accounting for dependence
0	94.66	96.66
30	58.33	96.33
60	45.66	95.33
90	41.00	95.66

Table 6: Coverage estimates, both incorrectly treating observations as independent and accounting for dependence, when the outcome initially follows a $Uniform[-1, 1]$ distribution.